DOCUMENT RESUME

ED 319 090 CS 507 166

AUTHOR Carrell, Thomas D.
TITLE Contributions of Fundamental Frequency, Formant Spacing, and Glottal Waveform to Talker Identification. Research on Speech Perception. Technical Report No. 5.
INSTITUTION Indiana Univ., Bloomington. Dept. of Psychology.
SPONS AGENCY National Institutes of Health (DHHS), Bethesda, Md.
PUB DATE 1 Aug 84
GRANT NS-12179-08
NOTE 160p.; Ph.D. Dissertation, Indiana University.
PUB TYPE Dissertations/Theses - Doctoral Dissertations (041) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS *Auditory Perception; Communication Research; Cues; Language Research; *Sex Differences; *Speech
IDENTIFIERS *Speech Perception; Speech Research; *Talker Identity; Waveform Analysis

ABSTRACT

This study investigated the contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification. The first two experiments focused on the effect of glottal waveform in the perception of talker identity. Subjects in the first experiment, 30 undergraduate students enrolled in an introductory psychology course, indicated whether stimuli were produced by a male or female. Results indicated that the crossover point of the identification function depended on whether the glottal waveform had been produced by a male or female talker. Subjects in the second experiment, 26 individuals chosen from a speech research laboratory's paid subject pool, were trained to identify a talker's gender. Results indicated that listeners could identify individual talkers on the basis of glottal waveform information alone. Experiments three (20 subjects) and four (13 subjects) assessed the relative contributions of glottal waveform, fundamental frequency, and formant spacing to the perception of talker identity. In both experiments, synthetic stimuli were constructed by copying the three cues of interest from tokens of words produced by natural talkers. Subjects were required to identify the talker. Results indicated that formant spacing and fundamental frequency were the primary sources of information in the speech waveform. Results also indicated that glottal waveform played only an indirect role in talker identification. (Thirty-seven figures and eight tables of data are included; 90 references, glottal waveforms, Fourier transforms of glottal waveforms, and one table of data are attached.) (RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED319090

RESEARCH ON SPEECH PERCEPTION

Technical Report No. 5
August 1, 1984

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana
47405

Supported by:

Department of Health and Human Services
U.S. Public Health Service

National Institutes of Health
Research Grant No. NS-12179-08

CS507166

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

THOMAS D. CARRELL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

RESEARCH ON SPEECH PERCEPTION

Technical Report No. 5

August 1, 1984

Speech Research Laboratory

Department of Psychology

Indiana University

Bloomington, Indiana 47405

Supported by:

Department of Health and Human Services

U.S. Public Health Service

National Institutes of Health

Research Grant No. NS-12179-08

Contributions of Fundamental Frequency, Formant Spacing, and Glottal Waveform to Talker Identification

Thomas D. Carrell

Submitted to the Faculty of the Graduate School
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
Indiana University
July, 1984

Table of Contents

Abstract ii

Acknowledgements.....iv

Chapter 1. Introduction.....1

Chapter 2. Analysis and Synthesis Techniques.....18

Chapter 3. Contributions of Glottal Waveform to the Perception of Talker Gender.....28

Chapter 4. Talker Identification as a Function of Glottal Waveform.....35

Chapter 5. Naturalness and Intelligibility of Synthetic Talkers.....47

Chapter 6. Talker Identification Using a Factorial Combination of Cues.....56

Chapter 7. Summary and Conclusions.....79

References 86

Appendix 1 91

Appendix 2 98

Appendix 3 105

Abstract

Contributions of Fundamental Frequency, Formant Spacing, and Glottal Waveform to the Perception of Talker Identity

Thomas D. Carrell

Indiana University

The contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification were examined in this investigation. Both fundamental frequency and formant spacing have been shown to be important in the perception of talker identity. Glottal waveform has been shown to differ substantially in the productions of different talkers. The overall goal of the present investigation was to examine each of these cues separately and in combination to determine their individual contributions to talker identity. This study also examined the way these individual attributes interacted with each other in controlling perception of a talker's gender.

The first two experiments focused on the effect of glottal waveform in the perception of talker identity. In the first experiment, glottal waveforms were extracted from talkers of different genders. A sensitive perceptual technique was used to assess the importance of glottal waveform in identification of the talkers gender by showing that the crossover point of the identification function depended on whether the glottal waveform had been produced by a male or female talker. These results were extended in Experiment 2 which showed that listeners could identify individual talkers on the basis of glottal waveform information alone.

Experiments 3 and 4 assesed the relative contributions of glottal waveform, fundamental frequency, and formant spacing to the perception of talker identity. In both experiments, synthetic stimuli were constructed by copying the three cues of interest from tokens of words produced by natural talkers. In Experiment 3, listeners rated the naturalness and intelligibility of these "synthetic talkers" on a seven point scale. The mean ratings were generally good, all falling within the upper half of the range of possible values, indicating that the synthesis methods used in these two experiments were adequate. We also found that naturalness and intelligibility ratings were not correlated, suggesting that these two measures assessed two different qualities of the speech stimuli. In Experiment 4, listeners were first trained

to recognize two male and two female talkers based on recordings of their natural voices. Synthetic stimuli were then constructed as in Experiment 3, but fundamental frequency, formant spacing, and glottal waveform from various speakers were combined in a factorial design. Listeners were required to identify the "talker" in each stimulus, although most stimuli contained cues for multiple talkers. The results indicated that formant spacing and fundamental frequency were the primary sources of information in the speech waveform that listeners use to recognize a talker's identity. Glottal waveform played only an indirect role in talker identification although it did directly influence ratings of naturalness of the synthetic speech. Therefore, while talker specific information is present in the glottal waveform and while it may be useful to listeners in certain perceptual tasks, the effects of fundamental frequency and formant spacing appear to play a much more important role in providing cues to talker identity in word length utterances. The results of this investigation have implications for improved speech synthesis methods and for the development of new techniques that can be used to identify individuals from the acoustic analysis of the speech waveform.

Acknowledgements

I have enjoyed the support of many people over the course of this investigation. I would first like to thank the members of my committee, David Pisoni, Linda Smith, Donald Robinson, and Robert Port for their efforts as well as their patience. I am especially indebted to David Pisoni who encouraged me to pursue the perception of talker identity and provided the background, advice, and facilities that made this dissertation possible.

Bob Bernacki contributed both programming support and creative input during all phases of this invesitgation. Jerry Forshee and David Link provided excellent system support.

Of course, I would never have been able to complete this work without the unfailing moral support of my wife June.

The research reported in this dissertation was supported by the National Institutes of Health, grant number NIH (NINCDS) NS-12179-08.

Chapter 1

Introduction

As the fields of human speech perception and machine speech recognition and synthesis have matured, relatively more attention has been directed towards some of the nonlinguistic or indexical properties of speech. This does not mean that all of the important phonetic and linguistic issues have been resolved; even some of the most fundamental issues in speech research continue to be controversial. What it does mean is that investigators working in these areas realize that some nonlinguistic characteristics of speech are important both to human listeners and to computer speech synthesis and recognition systems.

One of the most important of the indexical properties of speech is talker identity. In comparison to other attributes, the perceptual and acoustic characteristics of talker identity have been studied for a quite some time. No doubt this is due to the obvious forensic and military advantages to be gained by having the ability to make positive identifications on the basis of voice. More recently, however, other important reasons for studying acoustic cues to talker identity have become evident. One major concern is the desire to produce natural sounding synthetic speech that is acceptable to large numbers of listeners.

The need to improve synthetic speech is a natural outgrowth of the recent developments in producing relatively intelligible speech automatically and inexpensively. These developments, however, are still a long way from being perfect. Although listening tests with motivated subjects indicate that comprehension levels for some of the better synthetic speech is close to natural speech in paragraph length texts, more sensitive measures have shown large and reliable differences in perception between natural and synthetic speech when phonetic judgements are to be made or when competing tasks force listeners to devote less "attention" to the speech signal (see Pisoni, 1982; Luce, Feustel, & Pisoni, 1983).

The lack of natural phonetic quality has been one of the main problems in listener acceptance of synthetic speech. One of the main goals in developing speech synthesis systems has been to improve the output at this level. Progress here will come from specifying more detailed acoustic cues so that phonetic sequences are more responsive to their surrounding environment. However, in

spite of the problems that remain at this phonetic implementation level, it can be argued that speech synthesis techniques have reached intelligibility levels that will allow research on the naturalness of synthetic speech to be profitably conducted in parallel with further research on the segmental intelligibility of synthetic speech. The studies conducted in this investigation were primarily concerned with issues of naturalness as it is related to talker identification.

In general, given a certain minimum level of intelligibility, increases in the naturalness of synthetic speech lead to increases in the acceptability of the speech by the listener. A familiar example is natural speech that has been recorded at one speed and played back at another. Although this speech can be quite intelligible, it often sounds very unnatural to the listener. Another example is based on informal listening tests using time-varying-sinusoidal speech (see Remez, Rubin, Pisoni, & Carrell, 1981). Using this particularly unnatural sounding speech, we found that by altering the signal in a manner which increased naturalness but added no new phonetic information, the listener acceptability of the speech was much improved. Of course, the acceptability of a given type of synthetic speech depends to a large extent on the task in which the synthetic speech is used; for example, we would expect that in certain environments voice warning signals would need to sound quite natural to be taken seriously, whereas in other situations they might need to sound very odd to be discriminable and attract the attention of the listener.

While the naturalness of speech may or may not be directly dependent on the degree to which it sounds like a single talker, it is reasonable to assume that the converse dependency does hold true. That is, one way to make synthetic speech sound more natural is to make it sound like one particular talker. Therefore, not only will studying the acoustic cues which cause listeners to differentiate talkers lead directly to synthetic speech that can mimic particular talkers, but it should also lead, indirectly, to the identification of those dimensions of the signal that produce natural sounding speech.

A second reason for interest in the perception of talker identity is the fact that the ability to recognize talkers by voice is, in and of itself, an an important aspect of human perceptual behavior. At the most basic level, this is because humans are social animals and the recognition of an individual by voice is an important biological prerequisite necessary to maintain the survival of the species. There are also a number of other reasons to be interested in the human perception of talker differences. One is to improve experimental methodology by pursuing adequate stimulus definition. This is an especially

important problem in experiments where sensitive measures are being taken or long chains of inference are being drawn, as in procedures that use information processing methods to study psychological questions. The stimuli in such situations must be manipulated systematically on relevant dimensions since the results of such experiments and their interpretations are often influenced dramatically by relatively small changes in the stimulus quality.

Over the last two decades information processing approaches to modeling human cognitive processes have been quite successful. A large number of experiments falling into this classification have used speech input and output in order to study processes that are not directly observable. As these models become more sophisticated, very detailed attributes of the stimuli must also be modeled. One example is provided by the work of Drewnowski and Murdock (1980). In several short-term-memory experiments, these investigators demonstrated that memory for words presented in lists does not take place on a word-unit basis. Analysis of intrusion errors indicated that sub-word length features, such as number of syllables, syllabic stress pattern, identity of the stressed vowel, and identity of initial and final phoneme, were used in recall. When studies of memory begin to take details such as these into account, the stimuli must be defined very carefully along physical dimensions that are perceptually relevant to the subject. Although this particular study is not directly related to talker identification, it does demonstrate the level of stimulus description that is becoming necessary in studies of higher level processes and the cues used to identify a talker are clearly an important characteristic of the speech signal.

A number of experiments have demonstrated specifically that talker identity is a necessary component in models of human perception and memory. One example of such an experiment was based on a very simple recognition memory task (Craik & Kirsner, 1974). In this experiment, subjects recognized words faster and more accurately when the words were presented again in the same voice than when they were presented again in a different voice. What aspects of the stimulus led to this effect? Was it simply fundamental frequency differences, or did the listener have to recognize the stimulus as coming from a different talker? If so, then what cues does a listener use to identify talkers at this level? These questions have not yet been approached by researchers working in the mainstream of memory research and still remain to be answered by speech researchers.

Talker identity has also been shown to be an important factor in other experiments, ranging from the study of relatively low level-phenomena such as "Posner" type name-code matching (Cole, Coltheart, & Allard, 1974; Allard &

Henderson, 1976), to the study of relatively high-level phenomena such as the recall of sentences (Geiselman & Bellezza, 1976). In the first example, subjects were presented with pairs of spoken letter names in either same-voice or different-voice pairs. The reaction time results suggested a clearly defined model similar to (although not identical with) Posner's (1969) visual model. As in Posner's letter matching task in which visually identical letters (e.g. A-A) were identified as the same more quickly than were visually dissimilar letters (e.g. A-a), Cole et al. found that "same" judgements were faster for same-voiced letters than for different-voiced letters. However, unlike Posner's results in the visual modality, "different" judgements were also faster for same-voiced stimuli.

In the second example, subjects were asked to recall a list of 20 sentences and the identity and location (right or left) of the talkers. Some of the subjects were instructed beforehand that they would be asked to recall the identity of the talkers in addition to recalling the sentences, others were told that they would also be required to recall the location, and the remaining subjects were given no information indicating that they would be asked to recall anything other than the sentences. The results indicated that asking subjects to be prepared to recall the identity of a talker had no effect on the accuracy of sentence recall whereas asking them to be prepared to recall the location of sentence presentation reduced the accuracy of sentence recall. This finding was interpreted to mean that voice was automatically encoded along with the meaning of the sentence whereas location information had to be encoded separately. In both of these experiments, as in the Craik & Kirsner list learning experiment, the specific source of these effects was not identified. What aspects of the speech waveform are being encoded separately by the listener? Perhaps the same talker at different pitches would show the same effects? Whether talker identity is encoded holistically as a unit, or different nonlinguistic aspects of the speech signal are encoded independently, is simply not known at this time.

In order to design experiments that probe the mechanisms underlying phenomena such as these, a better understanding of the acoustic cues used in the perception of talker identity will be necessary. Similarly, in the immediate future, applied problems will require synthesized speech signals which adequately specify talker differences. Obviously, these two requirements -- better knowledge of cues and better speech synthesis -- will be approached hand in hand since progress in one area will be vital to progress in the other.

Closely related to the perceptual work described above are studies that have examined the differences in production between different talkers. One problem in examining the

literature on this topic is the finding that most studies that examine inter-subject variability in speech production do so with the intent of "normalizing it away" to more accurately decipher the underlying linguistic message. The results of this approach indicate substantial variation among subjects on a particular physical dimension but give no indication of the reliability or the nature of these variations. As background for our own work on this problem, the following paragraphs summarize some of the major differences in production between talkers.

Production Differences Between Individual Speakers

Numerous cues are available in the speech waveform which might reliably distinguish one talker from another. A large number of these have been studied over the last thirty years. Some are based on well-known anatomical and physiological differences in the vocal mechanisms of different talkers. Others are related to the dynamic or temporal aspects of speech which are assumed to be based on learned articulatory motor control processes rather than on structural constraints. The first set of differences, those related to the talker's vocal-tract anatomy, are based on the relative size and shape of the respiratory system, the larynx, and the superlaryngeal vocal-tract of different talkers. These differences appear most clearly in the relative frequencies of the formants and the power spectrum of the glottal source.

The origins of many talker differences can be accounted for within the Acoustic Theory of Speech Production (Fant, 1960). This model relates properties of the speech production mechanism to characteristics of the radiated speech waveform. The major assumption of this model is that the speech production system can be divided into two relatively independent components, a sound source that creates acoustic energy across a wide range of frequencies, and a time-varying filter or "transfer function" that shapes this energy by emphasizing certain frequencies and attenuating others. In normal voiced speech, the opening and closing of the vocal folds at regular intervals creates the source of energy; the superlaryngeal air passages above the glottis make up the filter. Figure 1.1 is a schematic midsagittal section of the adult human vocal tract showing the major components of the human speech production.

The regular vibration of the vocal folds produces a nearly periodic fundamental frequency with harmonics that gradually diminish in energy at higher frequencies. The fundamental frequency of this vibration is controlled by the size and tension of the vocal folds and is related to

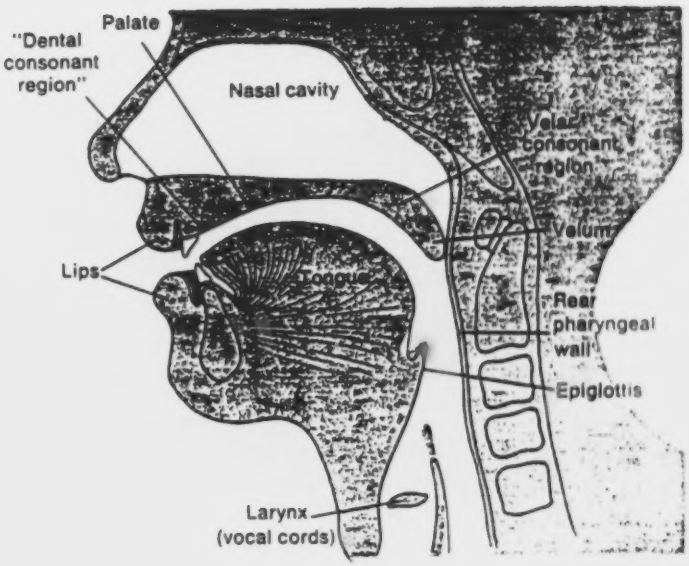


Figure 1.1 Schematic midsagittal section of adult vocal tract (from Lieberman, 1977).

perceived pitch. The relative levels of energy at different harmonics determine the perceived sound quality. For example, a male speaking in falsetto, where the glottal waveform is very sinusoidal sounds quite different than a female speaking with normal phonation at the same fundamental frequency. This is due, in part, to the fact that there is much less energy at higher harmonics with a falsetto compared to a normal source of phonation.

The filtering properties of the superlaryngeal vocal tract are determined both by the natural size and shape of these cavities and by the position and shape of articulators (such as the tongue and jaw). Since the latter characteristics are under the talker's conscious control, they are generally used for linguistic purposes whereas those aspects of the production mechanism that are relatively fixed are most often used to convey information about talker identification.

A schematic representation of the way the superlaryngeal cavities filter the source waveform is shown in a series of panels in Figure 1.2. Panel A shows the glottal waveform (the volume velocity at the glottis), and Panel D, below it, shows the energy spectrum of this waveform, that is, the energy level of each harmonic of the glottal waveform. Panel B shows the vocal tract shape and Panel E, below it shows the vocal tract transfer function, that is, the filter characteristics of the superlaryngeal tract. This transfer function is multiplied by the source function to determine the output waveform. Panel C shows the output of the system -- the radiated air pressure waveform measured at some specified distance in front of the lips. Panel F below it shows the short-term energy spectrum of this waveform. As an example of the effect of articulator position, Figure 1.3 shows the position of the tongue for the vowels /i/, /ɑ/, and /u/ and the resulting spectrum envelopes.

The formant patterns of vowels produced by different talkers have been studied in great detail over the years. In his classic monograph, Acoustic Phonetics, Joos (1948) was one of the first investigators to notice and discuss the marked differences in formant frequency patterns for the same vowels produced by different talkers. He was also one of the first to propose that some form of normalization must be carried out by the human listener to perceive physically different acoustic signals as the same vowel category.

These observations were further quantified by Peterson and Barney (1952) in their classic study on the formant structure of 10 vowels produced by 76 men, women, and children. Since that time, vowel formant frequencies have been found to be reliably correlated with a talker's size, sex, and identity. In fact, in a listening test in which

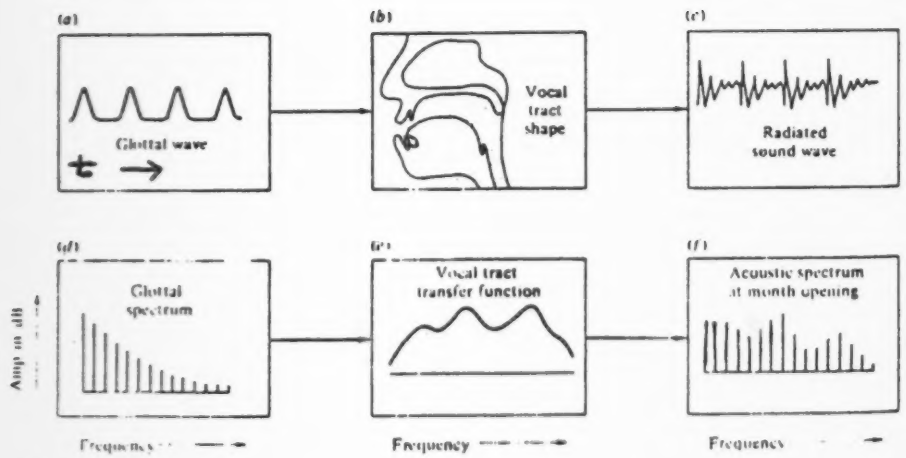


Figure 1.2. Components of the speech production system illustrating the Acoustic Theory of Speech Production (after Fant, 1970).

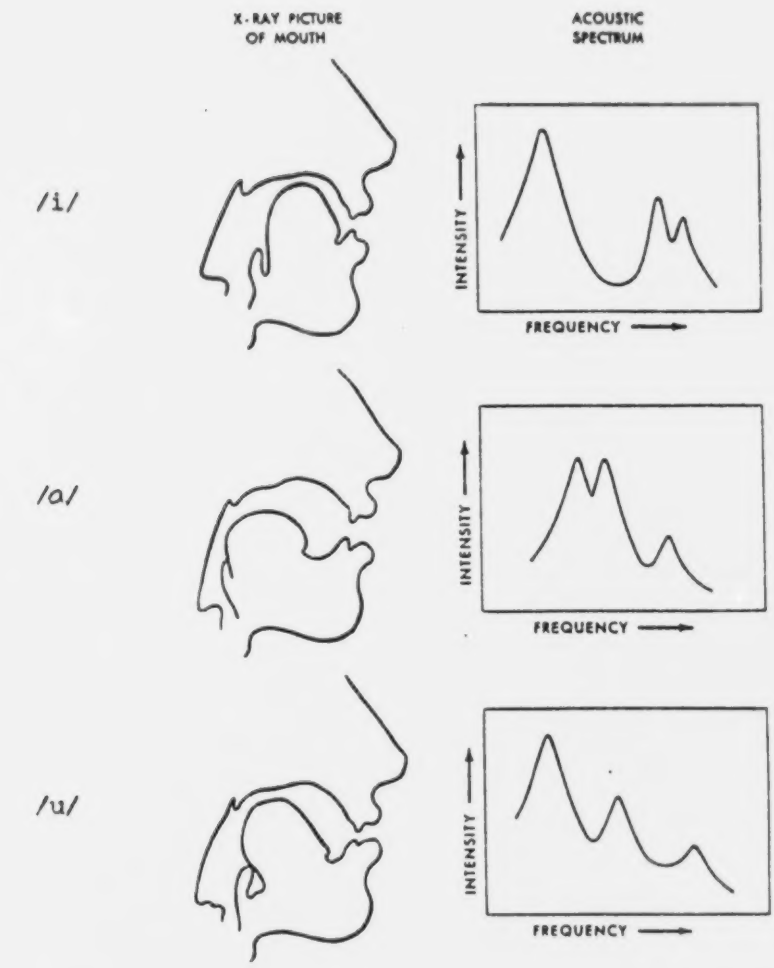


Figure 1.3. Position of the tongue and energy spectrum for /i/, /a/, and /u/.

subjects were required to identify 10 vowels and also identify whether the vowel was spoken by a male, female, or child, the talker identification performance was actually higher than the vowel identification performance (Lehiste & Meltzer, 1973).

Other investigators have found that formant bandwidth and overall vowel intensity differ across talkers. For example, Dunn (1961) found very large bandwidth differences between speakers as well as between vowels. And, Sharf (1966) found large differences between three speakers in two different measures of relative vowel intensity.

Another source of talker differences that is based, in part, on superlaryngeal anatomy is the short-term energy spectrum of fricatives. Several studies have shown that the spectra for the same fricative consonant were quite dissimilar across different talkers. In fact, the inter-talker differences in fricative spectra in a study by Hughes and Halle (1956) were so great that these authors were foiled in their attempts to graphically depict the average spectral shape of the fricatives that they examined. In a more recent study that examined a number of different cues to talker identity, Wolf (1969) not only claimed that talkers differed reliably in their fricative production, but also that a very effective procedure for capturing the talker specific fricative information was to first divide fricatives into four classes: single narrow peak, wide or double peak, no peak, and very low-frequency major peak. The classification number was then used in combination with other cues to predict talker identity. While this method provided talker specific information it was not compared directly with any of the other possible methods of analyzing the spectrum of fricatives.

The nasal cavities also exhibit wide anatomical differences between talkers, and since they play a significant role in the generation of certain speech sounds they, too, produce acoustic differences across talkers. A number of studies support this conclusion. Glenn and Kleiner (1968) argued that unlike much of the speech production apparatus, the nasal cavities are relatively fixed and therefore are specially suited to be the basis of systematic differences between talkers. This claim was supported by an algorithm that was able to identify single individuals from a group of 30 talkers (20 males and 10 females) with an accuracy of 93 percent correct.

In Wolf's (1969) study, differences in the nasal resonances between different talkers were measured from the spectra of the nasal consonants [n] and [m]. In combination with other cues, they were found to be reliable predictors of talker identity. All samples of speech in this experiment were recorded in one session, however, and no

research has been reported demonstrating that such differences are robust over long periods of time. Given the physiology of the nasal cavities and their sensitivity to factors such as disease, heat, humidity, and pollen count, it is unfortunate that such a study has not been performed yet.

One of the largest sources of talker differences based on anatomical factors is the glottal source function. In early work, Flanagan (1958) failed to find very large differences in the source spectra of two talkers. However, since that study was reported a larger number of talkers have been examined and substantial differences have been observed.

One of the most comprehensive studies in this area was conducted by Carr and Trill (1964). These investigators used inverse filtering on six sustained vowels produced by ten male talkers to examine glottal variations across vowel type and talker identity. In inverse filtering, the vocal tract resonances are estimated and a filter is designed which precisely cancels their effect. A speech signal passed through such a filter should result in a signal nearly equivalent to that present at the glottis. This analysis demonstrated large differences in glottal waveform shape between talkers as well as smaller but significant differences between vowels. Slopes of the source spectra ranged from -8 to -16 dB per octave and variations were found in the shapes of the spectra and the number and location of zeros in the spectra. These findings suggest that glottal waveform information may be a reliable acoustic correlate of individual talkers. In fact, Stevens (1972) has stated that, "The structure that is probably responsible for the greatest inter- and intra-speaker variability in speech is the larynx, and, in particular, the vocal cords." Of course, the intra-speaker variability would need to be small or predictable enough for the glottal waveshape to be a useful cue to talker identity. Unfortunately, little work has been done on examining the changes in the voicing source from day to day over relatively long periods of time.

Setting aside this problem, Wolf (1972) used a relatively indirect measure of the glottal source waveform to good advantage in his automatic speaker recognition algorithm. The measure he developed was simply the difference in amplitude between the first and the third formants in the vowel [u], expressed in dB and normalized by their frequency separation. Such an estimate of the source spectrum based on a vowel spectrum is quite crude and is obviously affected by a number of unrelated variables. In spite of this, Wolf's measure was strongly related to talker identity.

More recently, glottal waveform variation was carefully examined for five male and five female talkers (Monsen & Engebretson, 1977). The results of their work replicated and extended earlier studies that showed that the shapes of the glottal waveforms were quite different for different speakers. One finding of particular interest was that differences in waveshapes were particularly large for male and female talkers. Female talkers exhibited more sinusoidal-like glottal waveforms whereas males exhibited more asymmetric waveforms.

Across all speakers, several types of variation of glottal waveform were observed in the time domain upon visual inspection. First, there were differences in the asymmetry between the opening and closing portions of the glottal waveform. Second, there were greater and lesser degrees of "hump" in the opening portions of the male glottal waveforms. And, finally, the duration of a complete glottal closure varied substantially between talkers. These differences are revealed in the frequency domain in terms of the spectral tilt of the energy in the glottal source. Monsen and Engebretson (1977) found intensity drops ranging from 12 to 18 dB per octave depending on the speaker and the frequency range examined. In the case of male talkers, the spectrum fell off at a rate averaging slightly less than 12 dB per octave below 1200 Hz and slightly more than 12 dB per octave for frequencies above 1200 Hz. For female talkers, the glottal spectrum fell off at an average of slightly less than 15 dB per octave below 1200 Hz and at more than 15 dB per octave at frequencies above 1200 Hz. If we assume that these results would hold true across a larger group of talkers, it is clear that not only are there substantial variations in the glottal source among different talkers but there are also systematic differences between male and female talkers.

A determination of whether or not there were reliable production differences in glottal waveform was not answered in the Monsen and Engebretson study for two reasons. First, a large enough sample of talkers was not examined (although this was also a problem in drawing conclusions about male versus female differences). And second, the same talkers were not examined repeatedly as would be necessary to allow comparisons of inter- and intra-subject variability to be made. However, their results do indicate that such a study would be fruitful and that the glottal waveform may be an important component of talker identity.

With the exception of glottal waveshape, the differences in production that we surveyed up to this point have not made any special distinctions between male and female talkers. There are, however, some very systematic differences that exist between genders. The most apparent

is fundamental frequency. In general, the speech from a female talker is produced at about one octave higher than the speech from a male talker. However, when the fundamental frequency of male speech is simply doubled and all other parameters are left untouched, the resulting speech waveform does not give the perception of female speech. Subjective evaluations usually indicate that the speech was produced by a prepubescent male or that it was neither precisely male nor female in origin.

Another major difference between male and female speech is that there is, on the average, about a 15% shift in the frequencies of the formants. The exact shift depends on the particular vowel examined. This shift is due to the fact that the overall vocal tract length is greater for males than females. Moreover, the ratio of the size of the oral and the pharyngeal cavities differs between males and females (Fant, 1966). After puberty the average length of the oral cavity is 9.1 cm while the pharynx is 8.25 cm for males. In contrast, both the oral cavity and the pharynx average about 7.0 cm for females (Fant, 1973). Thus, the differences in perception between male and female vowel formants depend on the particular formant of the particular vowel being spoken. Fant (1966) has proposed the "k-factor" to describe the number by which one multiplies an average male formant frequency to determine the average female formant frequency for a given vowel. Although there are undoubtedly many other differences between male and female talkers, simply knowing the fundamental frequency and the k-factor relationships captures much of the perceived differences between male and female talkers in a very concise manner.

The second class of acoustic cues to talker differences is based on an examination of the dynamic or temporal aspects of the speech signal. Less systematic research has been conducted on this topic than was the case with static cues, but a handful of studies have demonstrated that the temporal aspects of speech are indeed an interesting area in which to look for cues to talker identity. Significant individual differences have been found in the durations of vowels, glides, and consonants as well as in diphthongiza tion and shape of the fundamental frequency contour. Atal (1972), for example, successfully based an automatic speaker recognition method on temporal variations in fundamental frequency. And, in another study, based on the speech of 10 males, Goldstein (1976) studied the speaker-identifying features of formant track information. Each of the vowels in her study was placed in the carrier, "Say b_d again." Goldstein found that the time varying structure of diphthongs, tense vowels, and retroflex sounds were systematically related to specific characteristics of the individual talker.

A general conclusion that can be drawn from each of the studies cited, using either static or dynamic cues to talker differences, is that substantial individual differences are present in nearly any acoustic measure one derives from the speech signal. Do any of these differences have perceptual consequences? Can some of these be used in talker discrimination and talker identification? And if so, which ones are used? Although there have not been as many studies examining the perception of individual differences as there are known differences between talkers, it appears that some factors have been found to be useful in discriminating between individual talkers.

Perception of Acoustic Correlates of Talker Differences

The Perception of Talker Gender. One dependent measure that is often used in examining acoustical correlates of talker differences is the accuracy of listeners' judgments of talker gender. The reason for this is that, in general, whatever inter-talker differences exist, they do so to a greater extent between talkers of different sexes. Therefore, many of the studies that have claimed to examine the acoustical characteristics of talker differences have actually examined differences related to talker gender.

A number of studies have shown that the k-factor is perceptually important to listeners who were asked to make judgements about the male or female quality of the vowels. One of the earliest studies used whispered vowels to reduce glottal source and fundamental frequency information (Schwartz & Rine, 1968). Five male and five female talkers whispered the vowels /i/ and /a/. Subjects were then asked to identify the talkers' gender when these vowels were presented in a random order. The response accuracy in this task was 97.5 percent which led the authors to conclude that, "the primary acoustic cue that underlies the [gender] distinction appears to be the upward frequency displacement of the resonance peaks in the female vowels."

Coleman (1971, 1976) has also conducted studies to determine the importance of formant spacing on talker gender identity. Talkers produced vowels with the aid of an artificial larynx. This device provided a constant mechanical glottal source so that the only difference between the talkers was the shape and movement of the superlaryngeal cavity and articulators. Under these conditions, the gender of the talker was correctly identified 88 percent of the time. However, Coleman also found that when the frequency of the artificial larynx was changed between male and female levels, a male F0 always

predicted a male response whereas a female F0 predicted a response that depended on the k-factor. It is clear from these findings that formant spacing information is very important in determining talker gender identity, although it can be overridden by a very low fundamental frequency.

Sato (1974) performed listening tests using specially developed synthetic stimuli and found that four factors -- the slope of the overall spectral envelope which is closely related to the source spectrum, the fundamental frequency, the formant spacings, and the formant bandwidths -- were all very important for the identification of talker gender. Unfortunately, these results and the details of the testing conditions were only summarized very briefly, and the relative importance of these four factors was not described. In any case, Sato's results suggest that several additional factors may need to be considered in male and female talker identification than just fundamental frequency and formant spacing.

In one of the most detailed studies of talker gender identification, Lass, Hughes, Bowyer, Waters, & Bourne (1976) examined the relative contributions of fundamental frequency and vocal-tract resonance in a simple vowel environment. Six vowels were spoken by ten male and ten female talkers in two conditions, voiced and whispered speech. A third condition was added by low-pass filtering the voiced stimuli at 255 Hz. The rationale for the stimulus selection was as follows: The voiced vowels were used to as a control to determine the absolute level of talker identification, the whispered vowels were used as a measure of the contribution of the superlaryngeal tract, and the low-pass filtered vowels were used as a measure of the contribution of the source function. Listeners performed the talker gender identification task at accuracy levels of 96% for the natural stimuli, 75% for the whispered stimuli, and 91% for the filtered stimuli. Since the whispered stimuli were assumed to contain mostly vocal tract resonance information and the low-pass filtered stimuli were assumed to contain mostly glottal source information, the authors concluded, in contrast to Coleman (1976), that the glottal information was a more important attribute for speaker gender identification than was the vocal tract resonance information. This result is not too surprising, however, because the fundamental frequency was retained in the low-pass filtered condition and the subject's task was to identify voices as male or female.

One final cue to gender identification has been found in the spectrum of fricatives. Ingemann (1968) demonstrated that listeners could identify the gender of both familiar and novel talkers on the basis of fricative information at levels ranging from chance for /θ/ to 90 percent correct for /h/. In general, it was found that as the place of

articulation moved more posterior in the tract, the subject's response accuracy improved.

The studies reviewed in this section demonstrate that the identification of talker gender is easily performed using a variety of cues contained in the speech waveform. Performance is close to ceiling when fundamental frequency or vocal-tract resonance information is retained, and is above chance from fricative information alone. Although dynamic and long-term cues to gender identification were not considered here, some evidence suggests that they are important as well (see, for example, Kramer, 1977).

The Perception of Talker Identity. Fewer studies have examined the acoustic components leading to changes in the perception of within-gender talker identity. Pollack, Pickett, and Sumby (1954) were the first to measure the effects of manipulating some physical attributes of the speech signal. In a study of speaker recognition in monosyllabic words produced by eight familiar talkers, they found that high-pass filtering the words at 1000 Hz or low-pass filtering them at 500 Hz left talker recognition performance above 80 percent accuracy. Performance for unfiltered words was about 92 percent. This result set the theme for those to follow: Nearly any acoustic information specifying talker identity is useful, but none, taken by itself, appears to be absolutely necessary.

This was certainly true, for example, in a more recent study carried out by LaRiviere, (1975) that examined two components of talker identity that were conceptually more meaningful than the two passbands used in the Pollack et al. study. One of the major concerns of LaRiviere's work was to examine the relative contributions of the source function and the superlaryngeal resonance of vocal tracts in the perception of talker identity. The listeners in this experiment were quite familiar with the talkers whose identities they were required to judge. Although the stimuli consisted of sentences, vowels in various forms, and fricatives, the vowel conditions were the most relevant to the present work and they will be examined in detail below.

Four vowels (/i/, /u/, /ae/, and /a/) were produced by eight male talkers (who were familiar to the subjects) using techniques identical to those of Lass et al. (1976) in their talker identification experiment. Conditions 1 and 2 were whispered and voiced vowels and Condition 3 was a low-pass filtered version of the voiced vowel. In this experiment, however, a 200 Hz filter cutoff frequency was used rather than 255 Hz (reflecting the fact that in this case all the speakers were male). The rationale for the stimulus construction was also similar: the voiced vowels were used

as a control to determine the absolute level of talker identification, the whispered vowels were used as a measure of the contribution of the superlaryngeal tract, and the low-pass filtered vowels were used as a measure of the contribution of the source function. The stimuli were presented to listeners in a random sequence and they were required to identify the speaker on each trial by circling the appropriate talker's initials on a response form.

Performance in the control (voiced) conditions which contained both glottal and superlaryngeal cues, was 40.2%, performance in the whispered conditions was 21.8%, and performance in the filtered conditions was 20.7%. Chance performance in this task was 12%. As would be expected, La Riviere's results indicate, first, that within-gender talker identification tasks are much more difficult than between-gender identification tasks, and second, that neither glottal nor superlaryngeal cues taken individually lead to accuracy levels as high as those of naturally spoken words in which both cues are combined.

It should be noted here that whispered speech does not necessarily eliminate all glottal source information nor does low-pass filtered speech eliminate all vocal-tract resonance information. To circumvent the methodological problems of using whispered speech, a number of experiments have used the Western Electric artificial larynx. For example, Coleman (1973) used this device to study talker discrimination in the complete absence of glottal source information (since all talkers used exactly the same glottal source). Unfortunately, this work cannot be directly compared to that of LaRiviere for a number of reasons. First, Coleman used a discrimination task rather than an identification task. Second, unfamiliar rather than familiar talkers were used. And, finally, 5-second segments of speech were used rather than isolated vowels. In any case, Coleman obtained a 90% discrimination accuracy level which indicated that glottal spectrum and fundamental frequency information were by no means essential for talker discrimination.

In reference to the second issue, the low-pass filtering problem, no relevant studies have been conducted which would eliminate the methodological problems of the filtering method of isolating glottal spectrum effects in a talker identification task. However, the fundamental frequency and fundamental frequency contours have been shown to be useful for such identification. Abberton and Fourcin (1978) reported a series of experiments in which talkers were to be identified only on the basis of their pitch contour in the sentence, "Hello! How are you?" Since the stimuli only contained source information, it can be presumed that the words making up the sentence could not be understood. The subjects were required to identify the

speaker from five possible choices. All the voices used were those of well known fellow classmates. The synthesis techniques used in this study allowed fundamental frequency and durational cues to be varied independently. The resulting stimuli therefore sounded "remarkably human in quality." The results indicated that either cue was sufficient for speaker identification at levels well above chance. The response accuracy was 74% for F0-based identifications and 53% for duration-based identifications (where chance was 20%). It is obvious from these findings that at least the dynamic aspects of the source signal are useful for talker identification.

In summary, the perceptual literature summarized above indicates that fundamental frequency and formant spacing are unquestionably two extremely important factors controlling the perception of talker identity. It is also clear that glottal waveform or source spectrum is an important acoustic correlate reflecting differences between talkers and may be useful in the perception of talker identity. One goal of the present investigation was to examine the contribution of glottal waveform, fundamental frequency, and formant structure as important cues to talker identity. A second goal was to determine how these three factors interact with one another in a variety of different identification tasks. Fortunately, the technology is now available to construct stimuli in which it is possible to manipulate these three acoustic cues independently from one another. Such stimuli can then be presented to subjects for identification. Before any experiments were conducted, however, the utterances of several male and female talkers were analyzed to extract the three acoustic parameters of interest: fundamental frequency, formant pattern, and glottal waveshape. These acoustic measurements were then used in the construction of word lists produced by synthetic "standardized talkers." That is, the words were synthetically generated but they were modeled after the physical characteristics of known talkers. This method of stimulus analysis and generation therefore allowed all acoustic parameters differentiating one talker from another to be held constant with the exception of the specific ones being tested.

Overview of Experiments

The first two experiments examined the role of glottal waveshape in the perception of talker differences. In the first experiment, we assumed that if glottal waveshape was an important attribute, it should be especially salient when extracted from talkers of different genders. In this experiment, a continuum of speech sounds was constructed ranging from male to female based on the formant ratio differences between men and women. We predicted that when these formants were excited by a male glottal source, listeners would judge the stimuli to be male further into the stimulus continuum than when the formants were driven by a female glottal waveform. The results of this experiment should indicate whether or not glottal waveform is perceptually important to listeners when they are required to identify talker gender. If the results are positive then it would be worthwhile testing whether or not glottal waveform is important in the within-gender perception of talker identity. If glottal waveform is not important, then there would be little reason to consider this any further as an important parameter controlling the perception of talker gender.

In the second experiment, listeners were trained to identify different talkers. The same listeners were then presented with only the glottal waveforms of these talkers and were required to make identification judgements. This experiment is the converse of the study performed earlier by Coleman, (1973) in which talkers were identified using only superlaryngeal cues. Positive results in this experiment would suggest that glottal source information alone can be used in the identification of talkers.

The next two experiments assessed the relative contributions of glottal waveform, fundamental frequency, and formant structure in the identification of talkers. In the third experiment, listeners were presented with both natural and synthetic tokens of a set of words. The natural set were simply digitized versions of a list of spoken words. The synthetic set were constructed using a substantially modified version of the Klatt software synthesizer. This is a digital simulation of a terminal analog formant based synthesizer containing a source function which models the laryngeal and turbulent sources in the human vocal tract and a filter which models the superlaryngeal characteristics of the vocal tract. The source section of the program was entirely rewritten to allow any arbitrary glottal waveshape to be used as the driving signal. On each trial, listeners heard both the

natural and synthetic versions of a word and were asked to rate the synthetic token on two different scales: (1) how intelligible was it compared to the natural token and (2) how similar did it sound to the voice quality of the original talker. Since the synthesizer (as modified) instantiates a specific model of the speech production system, the results of this experiment should demonstrate the ability of synthetic speech specified in terms of formant structure, fundamental frequency, and glottal waveshape to adequately specify talker differences.

In the fourth experiment, listeners were trained to identify a group of talkers from their utterances of a list of natural isolated words. A second set of stimuli was then presented and listeners were required to identify each word as belonging to one of the talkers learned in the first part of the experiment. This second set of stimuli consisted of synthetic tokens containing a factorial combination of the three parameters of interest: F0, glottal waveshape, and formant structure. The listeners in this experiment were also required to rate each stimulus item on scales of intelligibility and naturalness. The results of this experiment should provide information about the relative importance of each of the three cues manipulated.

In summary, the present set of experiments was designed to examine the contribution of fundamental frequency, glottal waveform, and formant spacing on the perception of talker identity and perceived naturalness in ways that have not been possible before. Until now, technical limitations have made it difficult to construct stimuli which would allow the examination of the independent and interactive effects of these acoustic characteristics of the speech signal. Furthermore, prior to the present investigation, different procedures have been used to examine different acoustic correlates of talker identity rendering a direct comparison between studies impossible. New synthesis techniques and a set of testing procedures were developed which allowed the effects of each of the three cues to be directly compared.

CHAPTER 2

Analysis and Synthesis Techniques

Several novel techniques were required for analysis and synthesis of the stimuli that were used in the present investigation. The general procedures and techniques will be described here rather than separately for each experiment since they all are quite similar.

Analysis of the Glottal Waveform

The glottal waveform is one of three acoustic correlates of talker identity that was examined and manipulated in this investigation. Historically, it has also been one of the more difficult components of the speech signal to isolate and analyze. This is primarily due to the physical inaccessibility of the larynx. Since the ability to accurately record the glottal waveform has been a serious problem over the past several decades, many different and often very creative techniques have been developed for this purpose. However, each method has its own drawbacks. The techniques range from rather drastic measures involving the removal of larynges from cadavers, to completely noninvasive measures, such as inverse filtering, which oftentimes only requires that speech be recorded from a good quality micro phone. Two of the more successful techniques which fall in between these two extremes are photographic and laryngographic measures.

In the case of the photographic techniques, highspeed motion picture photography has been used to measure the area of the glottal opening. The area is then converted to the volume velocity of air through the glottis, a process requiring a number of parameter estimates that are frequently difficult to determine precisely (Sondhi, 1975). In a much less invasive process, a laryngograph measures the electrical impedance between two electrodes placed on the surface of the neck. This method indirectly records the contact area of the vocal folds over time. A conversion to volume velocity is then required, resulting in problems similar to those of laryngeal photography. It is clear, however, from the few techniques that have been cited, that measuring the glottal waveform is indeed a difficult task; each of the techniques described so far has obvious practical limitations as well as problems of analysis and interpretation.

Another technique, inverse filtering, is probably the most widely used method for recording the glottal waveform at this time due to its relative ease of use and reasonable accuracy. In this technique, the spectral peaks corresponding to the resonances of the superlaryngeal tract during an utterance are determined by various analysis methods (such as linear prediction), and the speech waveform is then passed through a filter which is the inverse of this spectrum. The effect of such a process is to remove the resonances of the superlaryngeal passages thereby leaving the original glottal waveform intact. Unfortunately, the analysis methods for determining the resonances of the superlaryngeal tract are not perfect, and the filter may not precisely cancel the vocal tract resonances. The procedure that is generally followed in such a circumstance is to adjust the filter coefficients until the glottal waveform looks like it is "supposed" to. It is clear that subjectivity is the major drawback of inverse filtering.

The pseudoinfinite length tube. A number of years ago, Sohndi (1975) developed a device for recording the glottal waveform that has a number of very practical advantages and also provides a reasonably accurate representation of the glottal waveform. This device, called a reflectionless or pseudoinfinite length tube (PILT), was used for analysis of the glottal waveforms in this investigation. A complete description of its operation is provided in the two references just cited, but a brief summary will be given here.

The theory of operation of the PILT is based on extending the talker's superlaryngeal tract to an infinite length. Without a PILT, the human superlaryngeal tract (configured for a neutral vowel such as /ʌ/) may be modeled by a tube approximately 17 cm in length and 2.5 cm in diameter which is closed at one end (analogous to the glottis) and open at the other (analogous to the lips) as illustrated in Figure 2.1. An impedance mismatch at the lips creates a chamber with particular reflections and resonant frequencies. The resonant frequencies determine the formants of speech produced. When the PILT is added to the system, the impedance mismatch is eliminated at the lips without adding new reflections at the termination point of the PILT. This produces a system having virtually no resonances. The small microphone within the PILT will transduce the pressure waveform produced at the glottis into a voltage waveform which can be sampled by an analog-to-digital converter.

The PILT is very simple in construction and is diagrammed in Figure 2.2. A foam wedge at the termination

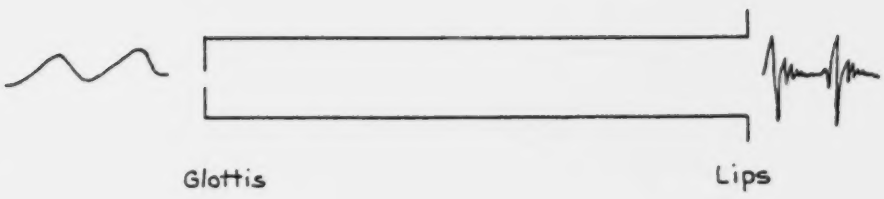


Figure 2.1. A simple acoustic tube model of the vocal tract in a neutral vowel configuration.

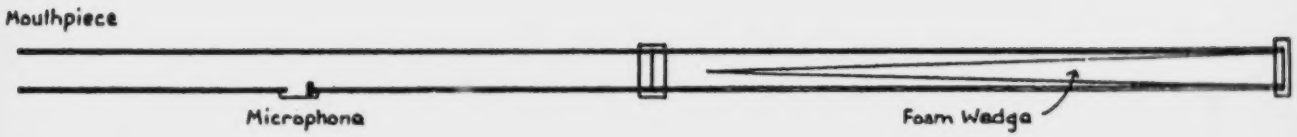


Figure 2.2. The pseudoinfinite length tube used for recording glottal waveforms. Note: The vertical scale is 2X the horizontal.

of this tube absorbs acoustic energy to prevent reflections. The air pressure measured by the microphone should be identical to that at the glottis with only a phase delay. The particular device used in the present investigation was 182 cm long with an inside diameter of 1.92 cm. The foam wedge was 82 cm in length.

In actual use, one end of this device is placed in the mouth of a talker with the lips tightly sealed. The talker adjusts his or her tongue to produce a neutral vowel. In order to obtain optimal performance, the oral cavity should be the same diameter as the tube and have no sharp bends or constrictions. Furthermore, flush contact must be made at the point the tube enters the mouth. In practice, of course, these ideal conditions are not always obtained, but when the vocal tract is configured to produce a neutral vowel, it provides an adequate approximation to the ideal. The effect of deviation from these assumptions will be discussed shortly. However, for the present it is sufficient to say that the deviations found in a typical recording session do allow for reasonable measurement of the glottal waveform.

The pseudoinfinite length tube was chosen for the present work for many of the same reasons that it was used by Monsen and Engebretson (1977) in their study of the glottal waveforms of 5 male and 5 female talkers. They stated that "it is physically uncomplicated and does not involve any discomfort or extensive preparation on the part of the human subject, it is highly noise-resistant and allows analysis of the glottal waveform in real time." Additionally, this technique was chosen because it nearly, although not completely, eliminated the subjectivity problems associated with the inverse filtering methods. This problem is not entirely eliminated with this method because the judgement of whether the coupling of the tube to the mouth was adequate or not was based on a real time display of the glottal waveform. If there was a pronounced hump in the opening phase of the waveform, the position of the pseudoinfinite length tube was readjusted and the talker attempted to change the vocal tract shape slightly. However, an opening phase hump has been assumed to be a real part of the glottal waveform of some male talkers by Monsen and Engebretson (1977) and is also well modeled by a laryngeal model developed by Ishizaka and Flanagan (1962). The PILT was therefore adjusted by the talker to reduce this hump as much as possible, because any inaccuracy in this method would be in the direction of increasing, rather than decreasing, the size of such a hump. In fact, none of the talkers in the present investigation exhibited this characteristic. Figure 2.3 shows the glottal waveform of the author as recorded by the PILT used in the present set of experiments.

Figure 2.3. Author's glottal waveform as recorded by the pseudo-infinite length tube. (A) Opening phase. (B) Closing phase. (C) Closed phase.

Sohndi (1975) claimed that the PILT was an excellent device for measuring the glottal waveform for a number of reasons. The primary one was that Fourier transforms of the output of this device showed no hint of formant peaks. Sohndi also offered the further observation (p. 231) that if the output of the PILT were given to an experimenter as the output of an inverse filter, "he would undoubtedly decide that the filter was accurately adjusted and accept the output as his best estimate of the glottal waveform!"

Such evidence, while putting this method in perspective with other methods of glottal waveform extraction, does not reveal the accuracy of the PILT on an absolute scale. The operation of the PILT was recently studied in depth; and now both its advantages and disadvantages are well understood (Hillman and Weinberg, 1981). In their study, a detailed model of the vocal tract (based on Stevens, Kasowski, & Fant, 1953) was combined with a model of the PILT. Known waveforms were used to drive the vocal tract model which was set to six different configurations: an ideal 17 cm uniform tube, and the vowels /ae/, /eh/, /^/, /u/, and /i/. The output of this model was coupled to the model of the PILT and the pressure waveform was measured within the PILT. When an ideal uniform tube vocal tract model was connected to the model of the PILT, the output waveform taken from the PILT was identical to the input at the source of the vocal tract model. When vowel configurations were connected to the PILT, however, the output waveform differed from the input in systematic ways. The difference was greater for the more extreme vowels, /u/ and /i/, than it was for the more neutral vowels, /ae/, /eh/, and /^/ but the difference was still substantial, even for the neutral vowels. The assumption of uniformity of the vocal tract was violated sufficiently by the vowel configurations to create low level standing waves which showed up as a slight ripple in the opening phase and closed phase, but not the closing phase of the glottal waveform. That is, the distortion appeared as a hump in the opening phase and made the closed phase appear shorter or nonexistent. Surprisingly, these features are not evident in Figure 2.3. In spite of the effects visible in the time domain, the amplitude spectrum of the output waveform was not shown to be different from that of the input.

In summary, the PILT method of measuring the glottal waveform was shown to be reasonably accurate in the time domain and very accurate in amplitude spectrum when the vocal tract was configured for a neutral vowel. Taken together, the PILT offered significant practical and theoretical advantages for measurement of the glottal waveform over the other methods examined.

The glottal waveforms recorded using the PILT for use in Experiments 2, 3, and 4 are shown in Appendix 1. Note that the characteristics that would be predicted by violating the uniform tube assumptions (opening phase hump and lack of a real closed phase) were not observed. Many earlier published waveforms based on the PILT technique (Monsen & Engebretson, 1977; Sohndi, 1975) did exhibit these problems. The reason that the present glottal waveforms appear to be more accurate is not presently known.

Analysis of Formant Frequencies

Formant frequencies were analyzed by two programs: SPECTRUM and SFVIEW. SPECTRUM is a general purpose spectral analysis program developed at the Speech Research Laboratory at Indiana University (Keweley-Port, 1979). This program was used to perform LPC analyses (Markel & Gray, 1976) on the speech signals used in this investigation. When a signal is analyzed, SPECTRUM creates a file for its own use that contains analysis information such as the smoothed spectrum, formant tracks, and reflection coefficients. As with many other formant peak picking algorithms, the one used in SPECTRUM leaves much to be desired. Invariably, hand editing of the formant tracks is required.

SFVIEW is an interactive program that was developed to make accurate formant tracking "by hand" relatively simple. It uses as input the smoothed spectrum taken from the SPECTRUM analysis file. Thus, the stimuli must be analyzed using SPECTRUM before the formant tracks are specified "by hand" with SFVIEW. In the present work, SPECTRUM was used to obtain smoothed spectra only; its formant tracking capabilities were not used. Each signal was analyzed with a frame size of 10 msec, a 25.6 msec Hamming window, and an LPC analysis requiring 12 coefficients. In some cases, the number of coefficients in the analysis was changed slightly to better discriminate the peaks.

SFVIEW simultaneously displayed two panels of graphic information on a DEC VT11 display screen. The top window showed a frequency-by-time plot of up to 5 formants, giving the impression of a schematized spectrogram. The bottom window showed an amplitude-by-intensity display of the smoothed spectrum of the utterance at a given point in time. This display is shown in Figure 2.4. A Summagraphics digitizing tablet was used by the experimenter to directly control the position of a triangle shaped cursor in the top window. However, the position of another cursor in the bottom window was specified only along the horizontal or frequency axis by the graphics tablet. The position of the cursor on the vertical (amplitude) axis was determined by the characteristics of the signal.

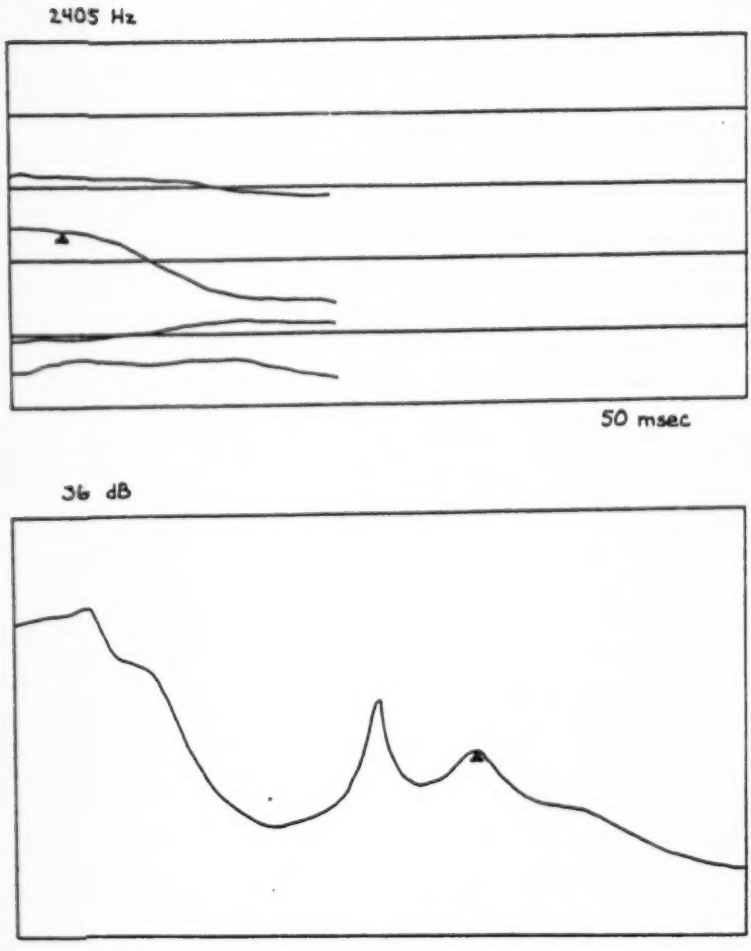


Figure 2.4. Display from SFVIEW, an interactive spectral analysis program.

For each signal to be analyzed, the name of the SPECTRUM analysis file was input, after which the VT11 immediately displayed the smoothed spectrum of the first frame (10 msec) of the utterance in the bottom window and a set of five default formant tracks in the top window. The mouse on the Summagraphics digitizing tablet directly controlled the position of the cursor on the top window. When the mouse was moved horizontally, the cursor moved along the time dimension in the top panel. When the mouse was moved vertically, again the cursor followed along the frequency dimension in the top panel. In addition to moving the cursor, two "odometer" type digital displays indicated the position in milliseconds in the file and the frequency in Hz. to which the cursor was pointing. In most cases, the mouse was moved at angles that were neither exactly horizontal or vertical, and naturally the cursor in the top window followed correctly and both odometers were updated simultaneously. Buttons on the mouse determined whether the cursor drew a line following its path or moved without drawing a line.

The bottom window displayed an amplitude section of the signal using a different set of coordinates than those used on the top; the horizontal axis represented frequency and the vertical axis represented amplitude. Time was not represented in this display. That is, the smoothed spectrum was shown for only a single point in time at once. As the experimenter moved the cursor in the top window, another cursor was moved in the bottom window, and the shape of the display was changed accordingly. Specifically, when the mouse was moved along the time dimension (horizontally), new spectral sections were displayed each time the cursor crossed into a new frame of the SPECTRUM analysis file: that is, every 10 msec. Moving the mouse rapidly in the horizontal direction gave the impression of a real time spectral display as the shape of the smoothed spectrum changed in the bottom window in time. Four consecutive frames from the bottom window are shown in Figure 2.5. Moving the mouse and, therefore, the cursor in the top window, vertically, moved a cursor in the bottom window horizontally along the frequency dimension. It was not possible (or meaningful) to directly move the cursor in the bottom window along the vertical, or amplitude, dimension since the cursor always fell on the line of the smoothed spectrum. Another "odometer" was also continuously updated in real time in the bottom window. This display showed the amplitude of the cursor position in dB. Thus, in the bottom window, the cursor could be moved in frequency by moving the mouse vertically; its position on the amplitude dimension was not under direct control of the experimenter. The height of the cursor in the bottom window was determined by the amplitude at the particular time and frequency pointed to by the mouse.

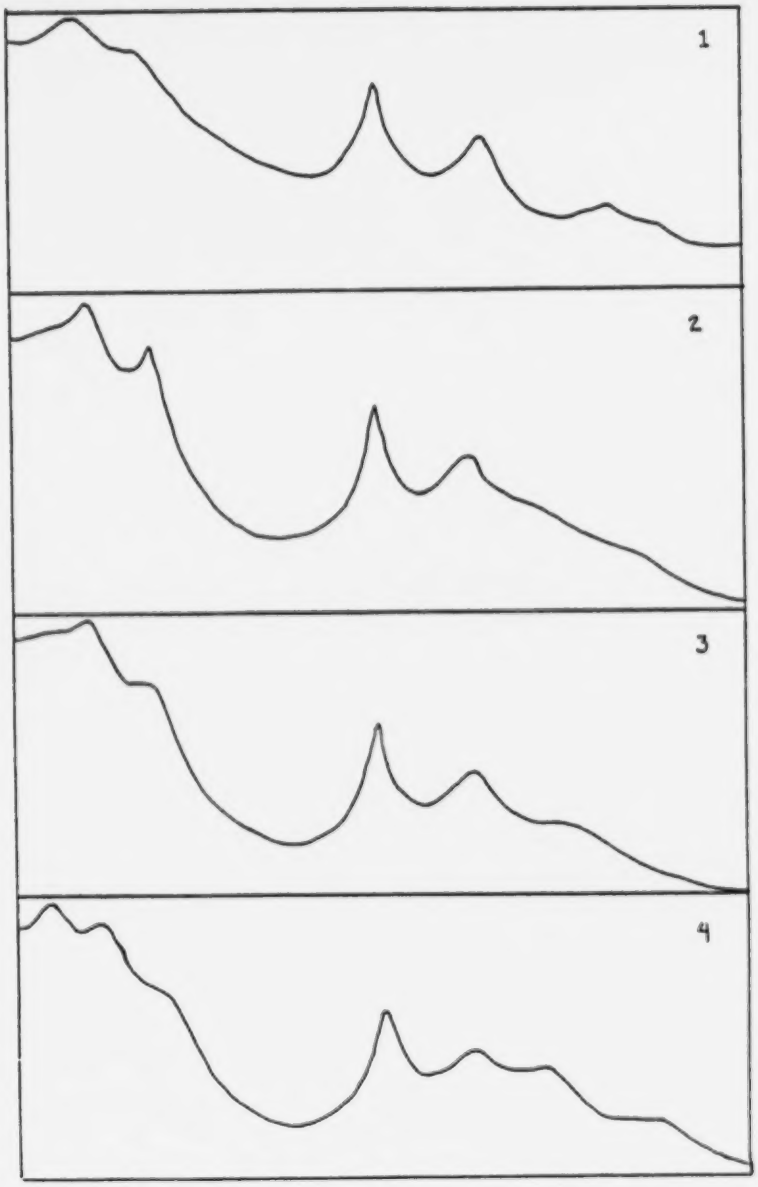


Figure 2.5. Four consecutive frames from the bottom window of SFVIEW.

In practice, one formant was traced in at a time, generally in numerical order. The cursor was positioned at the first frame on the amplitude peak for the first formant by positioning the mouse all the way to the left and then moving it vertically until the cursor was at the F1 peak in the bottom window. The cursor was next moved horizontally just far enough to reach the next 10 msec frame. At this point, the shape of the smoothed spectrum in the bottom window changed, and the cursor was again moved vertically to match the new peak. This process was continued across the length of the utterance with the cursor tracing the formant path on the upper window. Since it was possible to edit the formant tracings, generally a rough approximation was drawn first and then the mouse was used to make the final trace more accurate. After the first formant was input to the experimenter's satisfaction, the second through the fifth formant (or fourth in the case of female talkers) were similarly input. Although the process as described seems complex, in practice tracing in formants was simple but time consuming.

Why were such elaborate programs developed to aid formant tracking "by hand" rather than simply using automated formant tracking algorithms? The main reason is that the signals for the upcoming experiments were synthesized using a modified version of the Klatt software synthesizer (Klatt, 1980), a formant based synthesizer. The formants had to be very accurate for the synthesizer to produce both intelligible and talker-specific words. The talkers were not chosen for their "analyzability" nor did they repeat each utterance until it was easily analyzed by LPC curve fitting and formant tracking algorithms. The utterances of some talkers often had source spectra that were quite sinusoidal and therefore the source spectrum fell off rapidly enough so that no formants above the first formant were easily visible. Figure 2.6 shows a spectrogram of a particularly good example of this effect for the word "deed." Detailed examination of this token with SPECTRUM revealed that the second, third, and fourth formants were approximately 40 dB lower in amplitude than the first formant. It is very difficult to separate formants from noise when they are this weak. Many other problems exist in automatically measuring the formant patterns of a wide range of male and female talkers. The formant tracing technique used here remedies some of these problems by allowing a close and interactive examination of the signal.

The formant patterns traced in using SFVIEW were stored along with the formant amplitudes that were automatically measured in a new file which was directly compatible with our version of the Klatt synthesizer. This file allowed direct synthesis of the speech produced by real talkers as

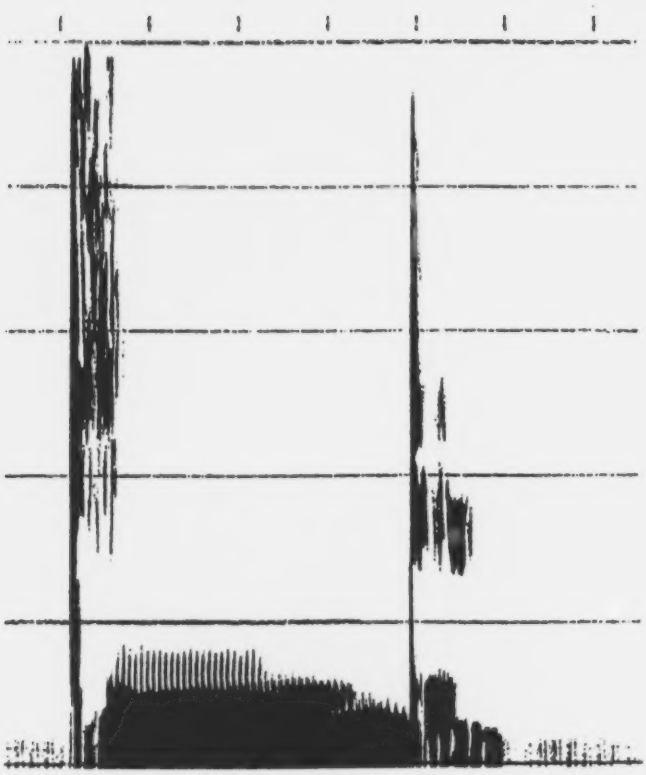


Figure 2.6 Wideband spectrogram of the word "deed" spoken by a male talker. Note the lack of energy in formants 2, 3, and 4.

well as experimental flexibility since parameters of interest could still be manipulated independently.

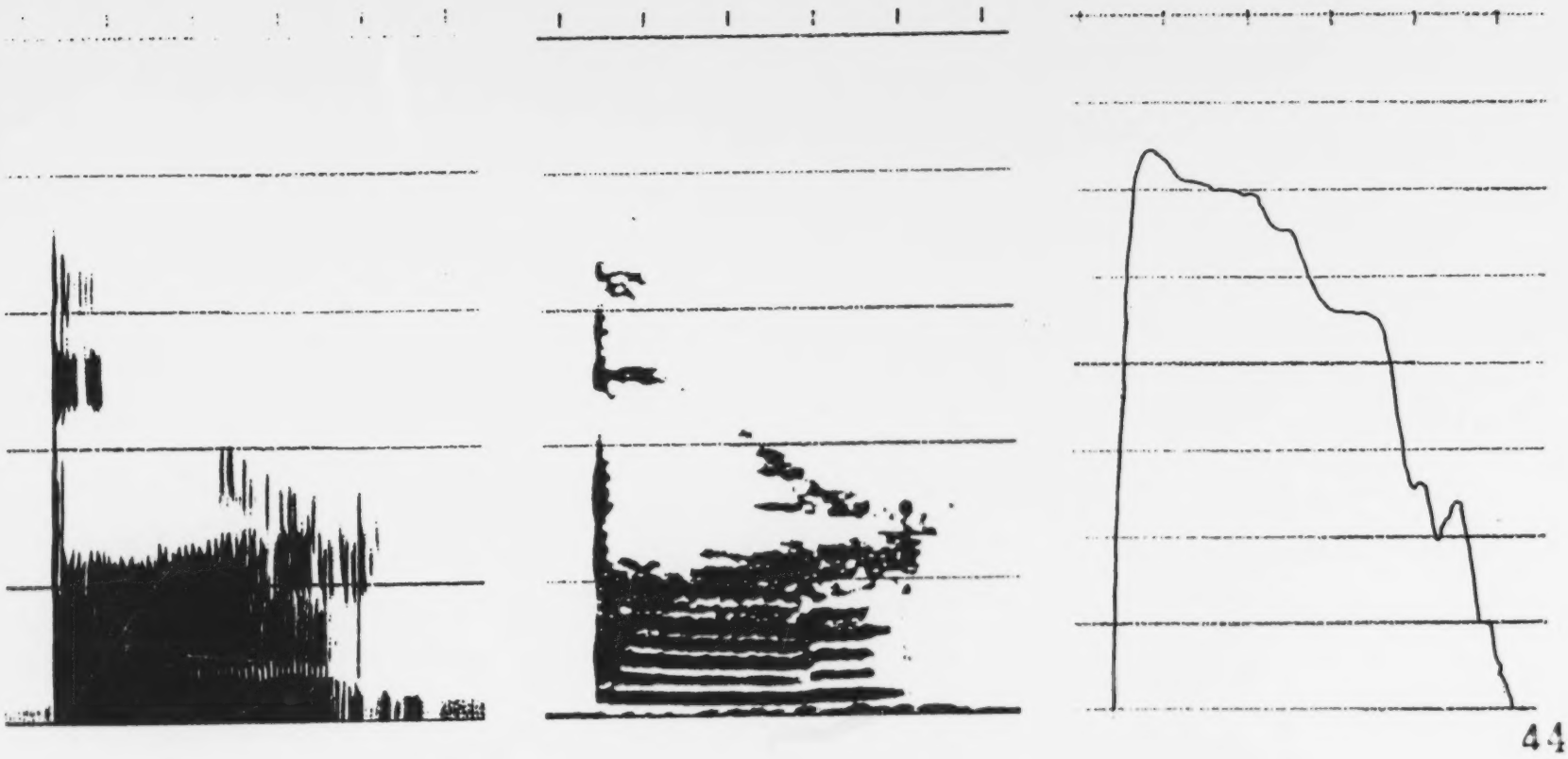
Analysis of Fundamental Frequencies

Fundamental frequencies were measured from a narrow band spectrogram of each utterance. Two types of spectrograms and an amplitude trace were made for each utterance using a Voice Identification Incorporated Series 8000 sound spectrograph. Generally, the tenth harmonic, as shown in the narrow band spectrogram (Panel B, Figure 2.7), was measured and the result was scaled down by a factor of 10; when this harmonic was absent other harmonics were used and scaled appropriately. Our version of the Klatt formant synthesizer allows graphical input and display of most synthesis parameters, and this facility was used to make and record the fundamental frequency measures from the spectrograms. The wideband spectrogram in Panel A was used to verify the results of the interactive formant tracking procedure, and the amplitude contour in Panel C was used to adjust the amplitudes of the voicing and frication sources.

The Klatt Software Synthesizer

Synthesis was accomplished with a version of the Klatt software synthesizer (Klatt, 1980) that has been adapted for use in the Speech Research Laboratory at Indiana University (Keweley-Port, 1978; Carrell & Kewley-Port, 1978; Bernacki, 1982). This is a very flexible formant-based digital synthesizer that allows specification of a large number of parameters every five milliseconds. The details of its structure and operation have been well reported elsewhere in the literature. However, its general functioning will be described here along with a description of a new extension of its capabilities that was developed specifically for the present investigation.

As summarized briefly in Chapter 1, the human speech production apparatus can be modeled as two independent components: a source and a filter. The Klatt software synthesizer is a digital implementation of this model with two basic sources and two configurations of filters. One source, corresponding to the glottal waveform of a human talker in phonated speech, consists of a pulse train which is passed through several filters to produce a naturally shaped glottal waveform. The parameters of the glottal filters can be adjusted to produce a wide variety of glottal waveforms. The second source is a pseudo-random number generator which is used to model the turbulence of the fricative source. Either of these sources can be fed into either of the two formant resonator configurations, parallel or cascade. In both configurations, digital resonators are



43

44

Figure 2.7. Wideband spectrogram, narrowband spectrogram, and amplitude contour of the word "bar" spoken by a male talker.

used to model the formants of the superlaryngeal vocal tract. In the cascade model, these resonators are connected in series; in the parallel model, they are connected in parallel. In general, the source is sent through the cascade branch for glottal (phonated or aspirated) sources, or through the parallel branch for superglottal (fricative) sources.

In the present investigation, three aspects of the speech signal were manipulated: (1) fundamental frequency, (2) formant spacing, and (3) glottal waveform. Of these, only the first two could be directly specified in the original version of the Klatt synthesizer. While it was possible to manipulate the glottal waveform source by using the three filters, and through trial and error arrive at a source spectrum close to that desired, it was not possible to directly enter any arbitrary glottal waveform to be used as the source for synthesis of phonated speech. Therefore, we modified the Klatt synthesizer to accept a file containing a glottal waveform to be used as the synthesis source.

Two programs were written to create glottal waveform (GLT) files. Each program created as output a file containing 512 points that specified one complete cycle. The first program allowed the experimenter to use the graphics tablet with mouse and the VT11 graphic display to simply trace in any glottal waveform. This was useful for experimenting with attributes of the signal that might be important for talker identification. This program was used to trace in previously published (Monsen and Engebretson, 1977) glottal waveforms for synthesis of the stimuli used in Experiment 1. The second program was used in conjunction with WAVES, a waveform editing program (see, Luce & Carrell, 1981), to convert a waveform taken from a PILT to the 512 point GLT format for the synthesizer. The waveform editing program was used to extract a single cycle from the 500 ms glottal waveforms digitized from the PILT.

Synthesis Procedures. In Experiments 3 and 4, the synthetic speech signals were based on ten word length utterances spoken by three male and three female talkers. The words were taken from the phonetically balanced PB list 1 (Egan, 1948). A rather involved procedure was used to synthesize each word as produced by each talker. The fundamental frequencies, formants, and glottal waveforms were first measured and inserted into the synthesis system as just described and a first attempt at synthesis was made. The output was then analyzed phonetically by the experimenter. If the word was not perceived as the intended one, for example, "pull" was heard when "pile" was intended, several things were checked. First, the natural model was reexamined. If the natural utterance sounded, phonetically, like the incorrect word, the synthetic word was considered

phonetically correct. This in fact happened several times in the stimulus set. Unfortunately, the decision was not always a clear one. In the natural context, there might be a number of redundant cues for one phoneme so that even though the formant spacings might be more like an /U/ than the intended /a^{y}/ (using the pull/pile example) and the natural word would generally be perceived as /a^{y}/ according to the context and the preponderance of cues, the synthetic word, due to the absence of redundant cues, might sound like /U/. If this was observed, then the word was synthesized according to the original formants rather than being "fixed." On the other hand, it was always possible that the formants were mistracked in the first place. This could be determined both by the experimenter's phonetic judgements and by LPC analysis of the synthetic version of the word and reanalysis of the model word. In this case, the synthesis parameters were corrected by returning to SFVIEW to examine peaks that may have been missed. If this did not solve the problem, the natural utterance was reanalyzed in SPECTRUM by realigning the time windows in relation to stimulus onset or by changing the number of poles in the analysis.

The analysis and synthesis methods described here usually produced synthetic speech that was intelligible and sounded like the natural speaker, although there was some variability in its success, as will be shown in Chapter 5.

CHAPTER 3

Contributions of Glottal Waveform to the Perception of Talker Gender

The present experiment was conducted to determine whether glottal waveform would have any measurable effect on the perception of talker gender. As noted in the Introduction, little evidence has been reported in the literature to address this question. Differences in production between male and female talkers suggest reliable and systematic differences. In the present study, both glottal waveform and formant spacing cues were manipulated to create a special set of synthetic stimuli that was used with a sensitive experimental procedure examine the existence and strength of this effect.

One of the most prominent differences in speech production between male and female talkers is the relative formant spacing or k-factor. Using synthesis techniques, it was possible to construct continua of vowels that range from male to female in voice quality by manipulating formant spacing. Furthermore, these manipulations can be performed without changing any other aspects of the signal. In this experiment a number of such continua were synthesized. Half of them used a male glottal waveform as the source (driving the formant resonators of a speech synthesizer) and half of them used a female glottal waveform to perform the same function. Listeners were asked to listen to each stimulus and indicate whether it was spoken by a male or a female talker. The dependent measure was the point along the k-factor stimulus continuum at which the subject crossed over from reporting a male to reporting a female talker. If glottal waveform plays a greater role in controlling the perception of talker gender than formant pattern, then all stimuli produced with a male glottal waveform should be perceived as a male talker and all stimuli produced with a female glottal waveform should be perceived as a female talker. On the other hand, if the glottal waveform is irrelevant to the perception of talker gender, then the crossover from male to female should always occur at the same point along the continuum regardless of the glottal source. In between these two extremes, the precise location of the crossover point may be used as an extremely sensitive measure of the relative contribution of glottal waveform to the perception of talker gender. For example, a stimulus produced with a male glottal source should cause responses of "Male" further towards the female end of the stimulus

continuum than the same stimulus produced with a female glottal source. We assumed that this methodology would be sensitive enough to determine whether or not glottal waveform was an important attribute used by listeners in perceiving a talker's gender.

Method

Subjects. Thirty Indiana University undergraduate students served as subjects in this experiment in partial fulfillment of an introductory psychology course require ment. None of the subjects reported any history of speech or hearing disorder and none had participated in any other experiments which used synthetic speech. All subjects were right-handed native speakers of English.

Stimuli. Six stimulus continua were synthesized for this experiment. Perceptually, each continuum ranged from a male to a female talker in six steps for a total of 36 stimuli. Half of the continua were synthesized with a male glottal source and the other half were synthesized with a female glottal source. Each of the three within-gender stimulus continua consisted of signals perceived as the vowels /i/, /ɑ/, and /u/.

The formant spacings for the continuua were determined in the following manner. For each vowel, two sets of formants were chosen, male and female. These values were obtained from the Peterson and Barney (1952) measurements of the formants of men and women. The value for the three steps between "male" and "female" on the continuum were linearly interpolated between the values of the endpoint stimuli. Based on the results of an earlier pilot study, an extra step was also created below the prototype male level, to help equate the number of male and female responses. The values for each of the stimuli are shown in Table 3.1.

Previous measurement studies have shown little overlap between the fundamental frequency distributions of male and female talkers (Fant, 1973, pp. 41). All stimuli were synthesized with a fundamental frequency which began at 180 Hz and dropped to 170 Hz over its 500 msec duration. This value was chosen because it falls into the ambiguous region between these two distributions. Pilot studies showed that depending on other stimulus parameters, vowels presented at this fundamental frequency could be heard as either a male or a female talker.

Two different glottal waveforms, one male and one female, were used as source functions in the synthesis of these stimuli. Both waveforms were taken directly from the glottal waveforms of two particular talkers (1M25 and 4F29) published by Monsen and Engebretson (1977). A special

Table 3.1

Formant Frequencies for all Stimuli -- Experiment 1

	k-factor	F1	F2	F3	F4	F5
	1	260	2164	2934	3075	3610
	2	270	2290	3010	3300	3750
	3	280	2416	3085	3525	3891
/i/	4	290	2542	3161	3750	4031
	5	300	2668	3236	3975	4172
	6	310	2794	3311	4200	4312
	1	700	1057	2348	3125	3610
	2	730	1090	2440	3300	3750
	3	760	1123	2532	3475	3891
/ɑ/	4	790	1156	2623	3650	4031
	5	820	1188	2715	3825	4172
	6	850	1221	2806	4000	4312
	1	283	850	2134	3150	3610
	2	300	870	2240	3300	3750
	3	317	890	2347	3450	3891
/u/	4	335	909	2453	3600	4031
	5	352	929	2560	3750	4172
	6	369	948	2666	3900	4312

version of the Klatt synthesizer (described in Chapter 2) was developed in order to allow graphic input of the desired source functions. In the standard version of the synthesizer, several parameters are used to specify the source function, and, while this makes it possible to create a wide variety of inputs to the formant resonators, there are important limitations in the type of waveforms that can be produced. These parameters control three digital filters which modify a pulse train to approximate a natural glottal waveform. Unfortunately, arbitrary waveforms, which might contain a number of poles and zeros in the spectral domain, cannot be created using this method. Since glottal zeros, for example, may be potentially useful in identifying talkers, such a limitation in synthesis could be constraining. Another advantage of the graphic method of specifying the source function is that it is a convenient common language which allows source waveforms created by many diverse methods, programs, and laboratories to be easily input into the synthesizer for further analysis and experimentation.

After the glottal waveforms were traced in they were stored in a 512 point normalized form. This waveform was sampled at appropriate intervals depending on the fundamental frequency desired. Each test stimulus was an isolated vowel, 500 msecs in duration, with 20 msec amplitude onset and offset ramps.

Apparatus. The experiment was conducted in real time under the control of a PDP-11/34 computer system that ran an experiment control program designed for this particular procedure. All stimuli were output at a rate of 10 KHz and were filtered with a very steep low-pass filter at 4.8 KHz (see Klatt, 1980 pp 990). Stimuli were presented to subjects over matched and calibrated TDH-39 headphones. The stimuli were presented at 80 dB SPL.

The experimental sessions were conducted in a sound attenuated subject testing room with six booths. Each booth contained a response box with two buttons, one labelled "Male" and the other labelled "Female." A warning lamp was centered at the top of the box.

Procedure. The subjects' task in this experiment was quite simple. On each trial a stimulus was presented and the subject was required to indicate whether the voice was produced by a male or female talker. The trial sequence proceeded as follows. At the beginning of each trial, the warning lamp on the response box was illuminated for 250 msec, indicating to the subject that a stimulus would follow immediately. Then, 500 msec later, a vowel was presented over the subject's headphones. At this point each subject had, 3.5 seconds to respond. If all six subjects responded before the end of the 3 second response interval, the next

trial was initiated immediately. The stimulus presentation order and each of the subjects' responses were recorded for later analysis.

Over the course of the experimental session, each of the 36 different stimuli (3 vowels by 2 genders by 6 k-factors) was presented 9 times each for a total of 324 trials.

Results

As expected, the glottal source function had a significant effect on subjects' responses of male or female for two of the three vowels. Both the /i/ and /u/ stimuli showed a crossover shift whereas /ɑ/ showed no shift. These differences are illustrated in Figure 3.1. The ordinate in this graph is the mean crossover point along the k-factor continuum from 1 (male) to 6 (female). Results based on stimuli created with a male glottal waveform are shown with solid bars, those from stimuli constructed with a female glottal waveform are shown with striped bars for each of the vowels /i/, /ɑ/, and /u/. The crossover for each subject was determined by fitting a logistic function to their data. The point at which 50% of the responses would be "Male" and 50% of the responses would be "Female" was considered the perceptual crossover. An example of this function is shown in Figure 3.2. The ordinate is the proportion of "Male" responses, the abscissa is the k-factor level.

Although a number of simpler methods have been used to calculate crossover points, the logistic function was chosen for a number of reasons. Most importantly, the logistic function is a reasonable model of the assumed underlying response distribution. This is due to the fact that it is an excellent approximation to the normal ogive which has frequently been used as an underlying representation of psychometric functions (and the logistic function is substantially easier to calculate than the cumulative normal). The data collected in the present experiment can be viewed as a psychometric function since it is the probability of making a binary decision, male or female, as a function of the level on a physical continuum, k-factor. There are also practical benefits in using this method above others. For example, one of the simpler methods of calculating crossover is to find the two points that lie on opposite sides of the 50% boundary and then linearly interpolate the level of k-factor for the 50% boundary. This method fails to take into account the possibility that, due to noise, the data may contain more than one crossover point, that is, the data may not be strictly monotonic. Should one take the crossover with the greatest slope? Average the crossovers? And, how would one calculate slope in either of these cases? The logistic function provides a

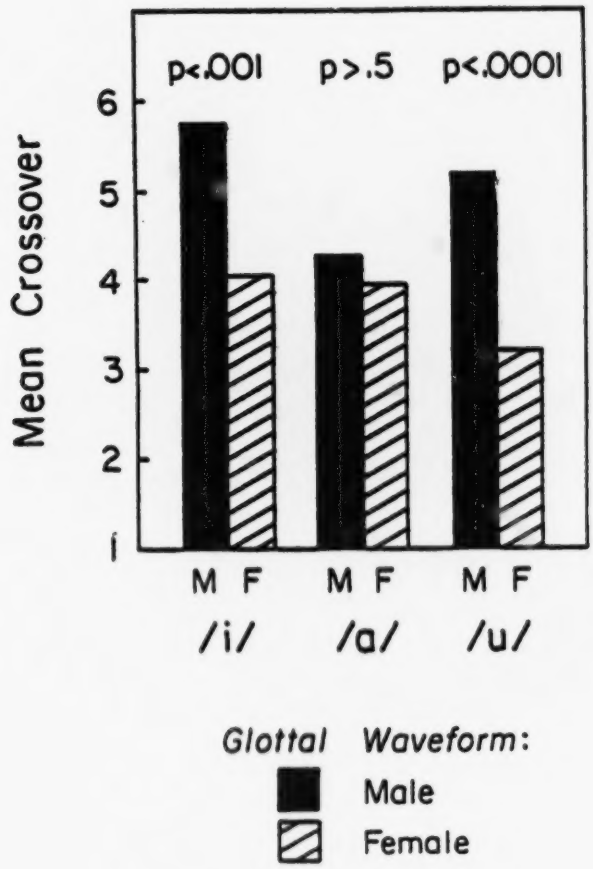


Figure 3.1. Mean crossovers from responses of "Male" to "Female" along the k-factor continuum for all stimuli.

theoretically motivated solution to such problems. The logistic functions used in the present analysis were calculated to be best fitting, in the least squares sense, based on a method provided in Neter and Wasserman (1974).

The best fitting logistic functions for the vowels /i/, /ɑ/, and /u/ are shown in Figures 3.2, 3.3, and 3.4, respectively. In the case of /i/, the crossover point for stimuli synthesized with the male glottal waveform clearly falls much later along the k-factor continuum, at 5.21, than does the crossover using stimuli based on a female glottal waveform, at 4.03 [$t(28) = 3.89$, $p < .001$]. The vowel /ɑ/, on the other hand, showed no such effect. In this case, the crossovers were nearly identical with values of 4.26 for the male condition and 3.96 for the female condition [$t(28) = .55$, n.s.]. The final vowel condition, /u/, showed the strongest effect of glottal source with a fitted male cross over of 5.64 and a fitted female crossover of -.21 [$t(28) = 4.84$, $p < .0001$]. The negative crossover reflects the fact that, regardless of the k-factor, the probability of labeling an /u/ synthesized with a female glottal waveform as male was less than 50%. The functions plotted in Figures 3.2 to 3.4 were averaged across all subjects; however, the statistical tests were performed on the crossover data from logistic functions fitted to the individual subject's responses.

A more direct way of examining the same data is to average the response probabilities for each of the stimuli. Histograms of this analysis are shown in Figures 3.5, 3.6, and 3.7. Here it can be seen that the k-factor manipulation produced continua ranging, perceptually, from male to female as expected. This was supported by the results of a two-way (glottal by formant) analysis of variance. All three vowels showed a main effect of formant, $F(5,145) = 85.6$, $p < .0001$; $F(5,145) = 47.6$, $p < .0001$; and $F(5,145) = 28.9$, $p < .0001$ for /i/, /ɑ/, and /u/ respectively. This analysis supports the same conclusions on the effects of glottal waveform as the earlier crossover analysis provided. For example, the response data from vowels /i/ and /u/ show a strong effect of glottal source on identification probabilities [$F(1,29) = 5.4$, $p < .03$, for /i/, and $F(1,29) = 42.2$, $p < .0001$ for /u/]. And, as in the crossover data, no main effect of glottal waveform for the /ɑ/ vowel was observed [$F(1,29) = .94$, n.s.].

The analysis of variance on the responses also revealed an interaction between glottal waveform and k-factor that was not apparent in the crossover analysis. Furthermore, this interaction was found for all three vowels [$F(5,145) = 10.9$, $p < .0001$; $F(5,145) = 36.1$, $p < .0001$; and $F(5,145) = 3.64$, $p < .005$ for /i/, /ɑ/, and /u/, respectively]. Visual inspection of the /i/ condition shown in Figure 3.5 reveals that this result was due to the fact that the glottal source

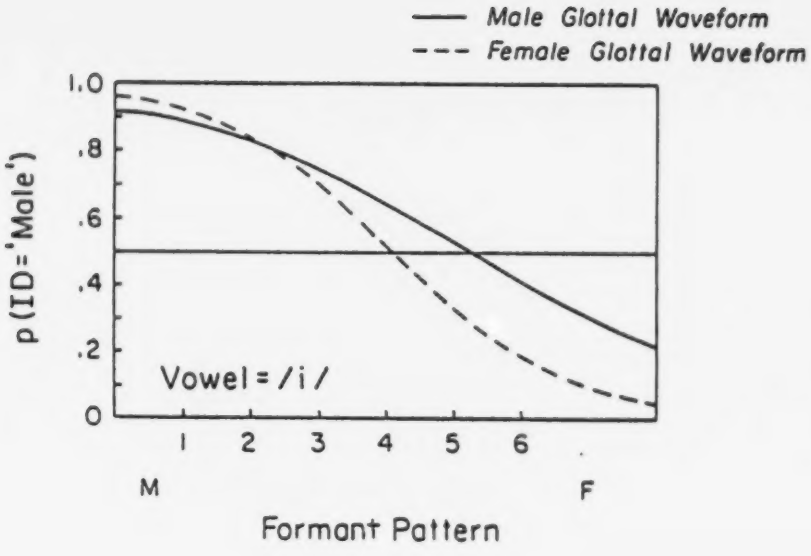


Figure 3.2. Best fitting logistic functions for /i/ stimuli synthesized with a male glottal waveform (solid line) and a female glottal waveform (dashed line).

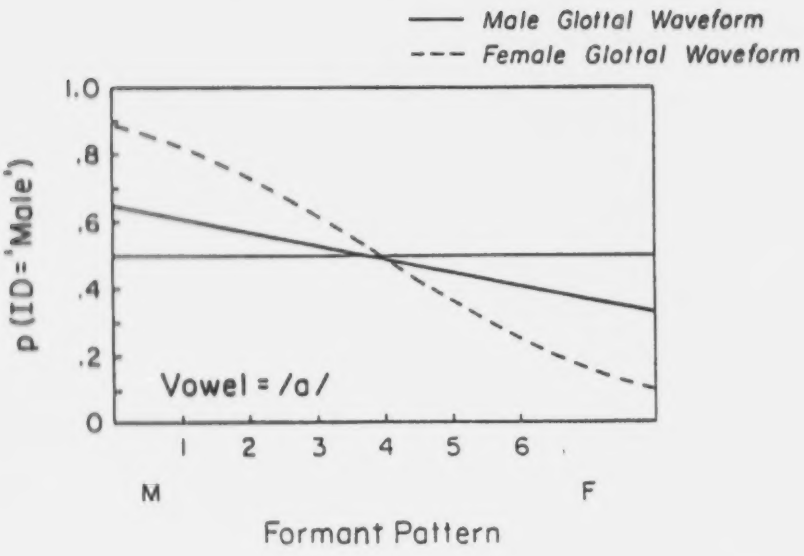


Figure 3.3. Best fitting logistic functions for /ɑ/ stimuli synthesized with a male glottal waveform (solid line) and a female glottal waveform (dashed line).

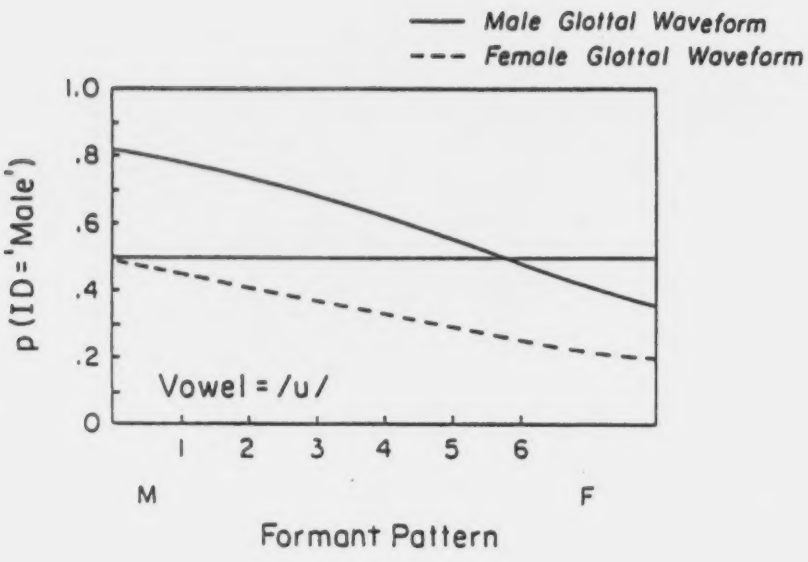


Figure 3.4. Best fitting logistic functions for /u/ stimuli synthesized with a male glottal waveform (solid line) and a female glottal waveform (dashed line).

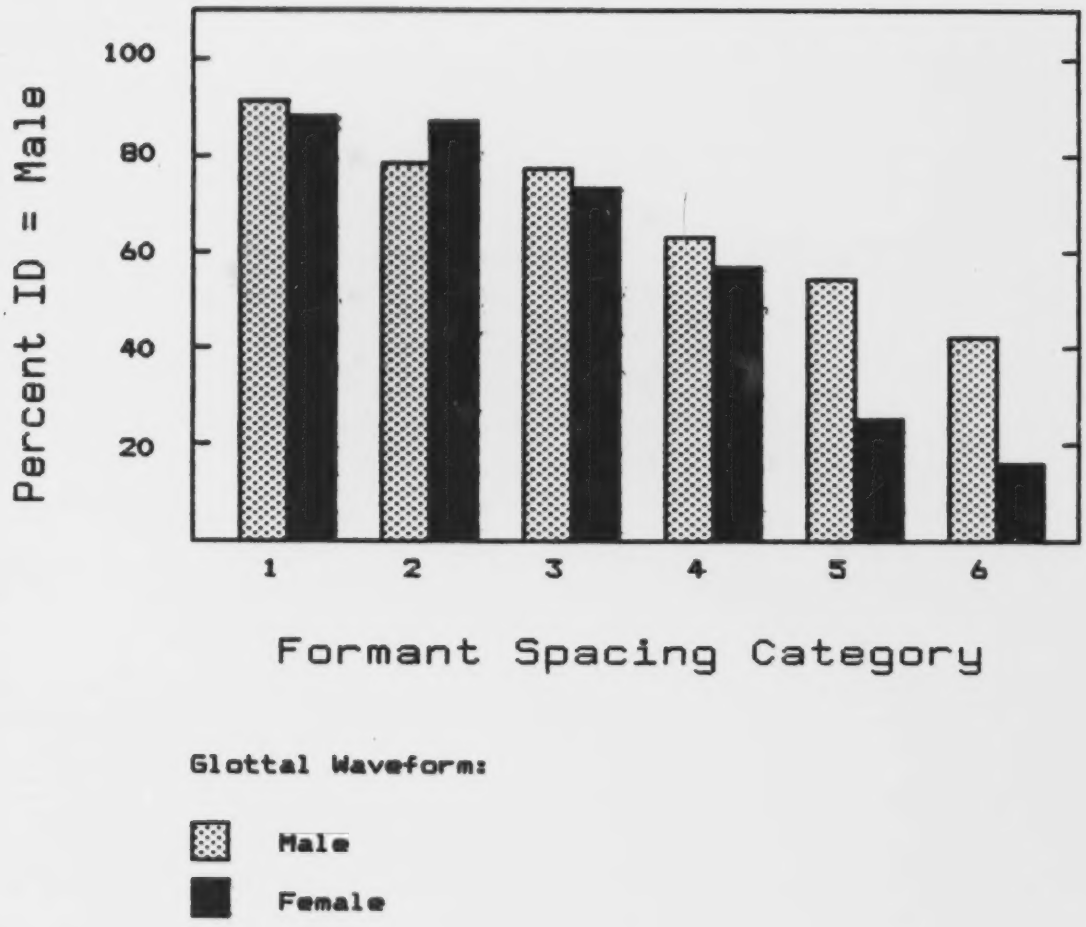


Figure 3.5. Mean percentage of "Male" responses along the six point male-to-female k-factor continuum for /i/ stimuli. The left bar of each pair was synthesized with the male glottal waveform and the right with the female glottal waveform.

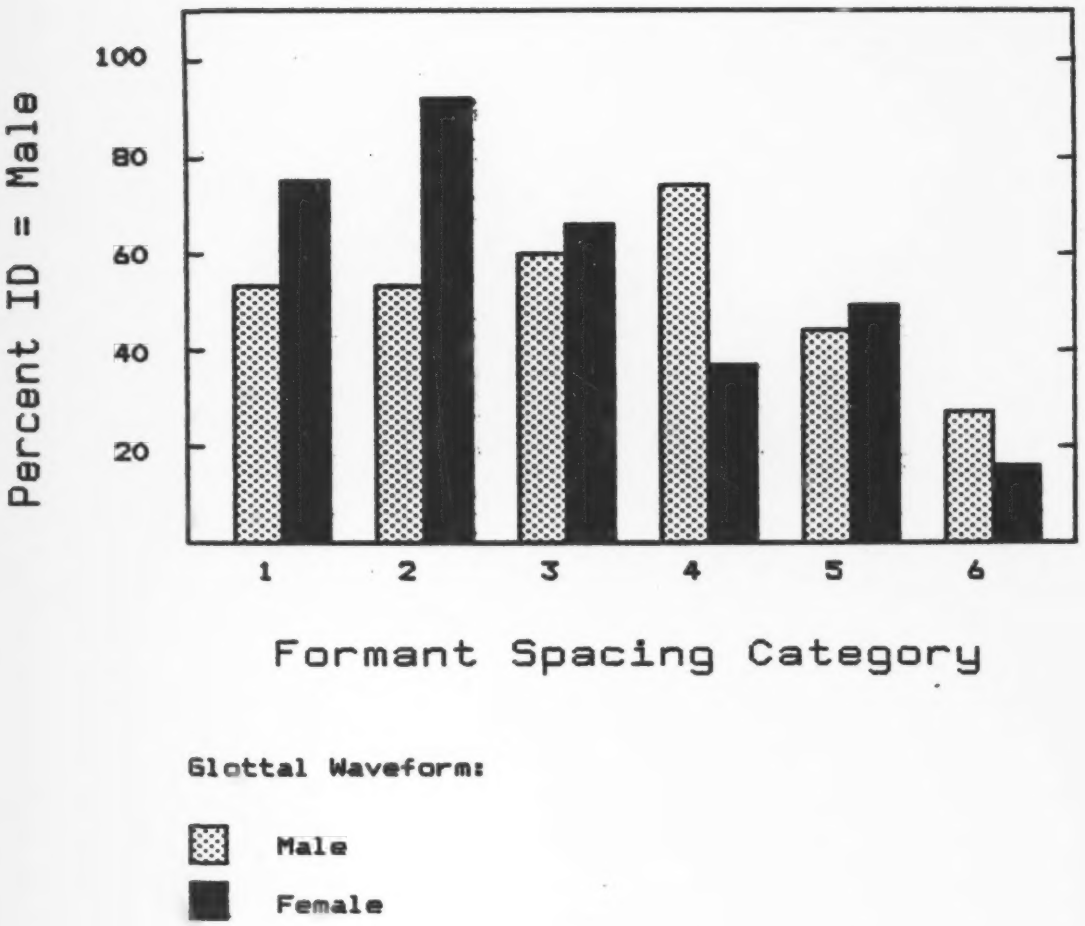


Figure 3.6. Mean percentage of "Male" responses along the six point male-to-female k-factor continuum for /ɑ/ stimuli. The left bar of each pair was synthesized with the male glottal waveform and the right with the female glottal waveform.

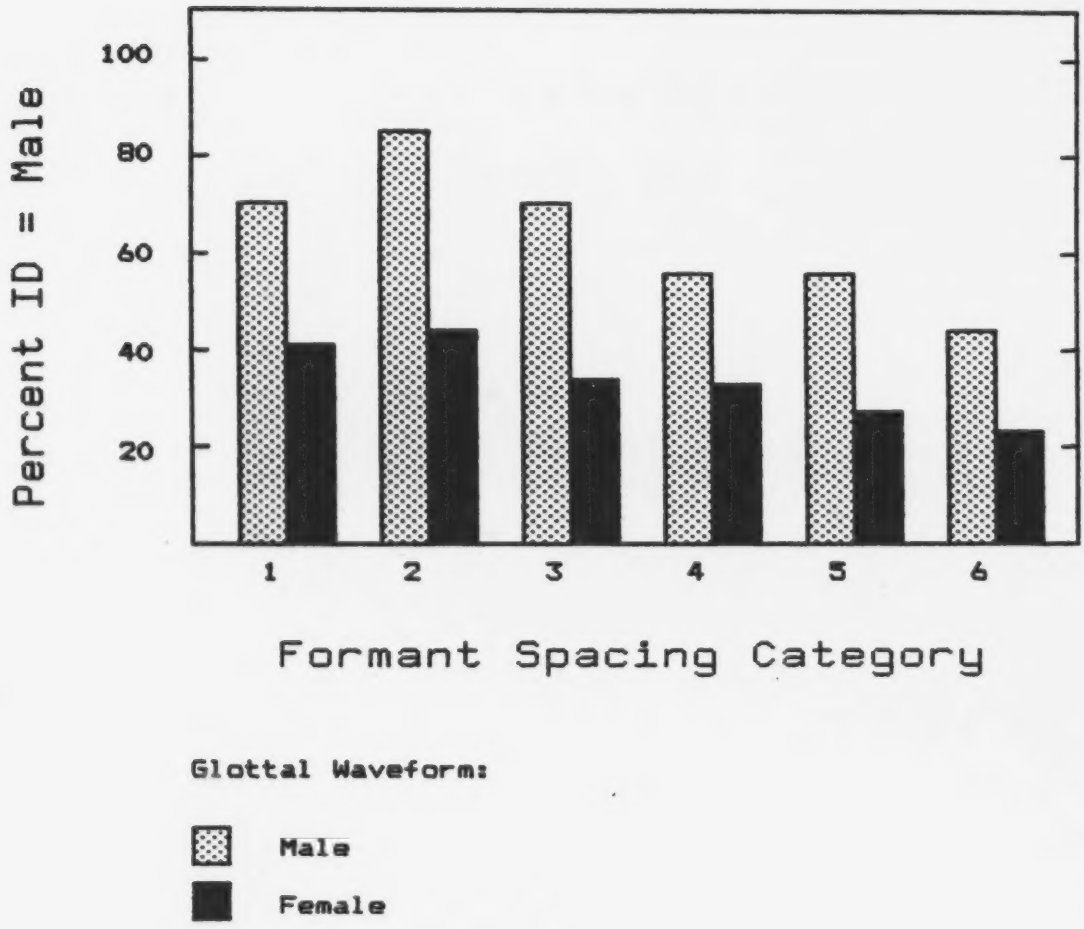


Figure 3.7. Mean percentage of "Male" responses along the six point male-to-female k-factor continuum for /u/ stimuli. The left bar of each pair was synthesized with the male glottal waveform and the right with the female glottal waveform.

only had its major effect at the female end of the k-factor continuum. The results are less interpretable in the other vowel conditions. The strength of the effect of the glottal source therefore depends on the specific formant patterns with which it was combined.

Discussion

The results of this experiment demonstrate that glottal waveform plays an important role in the perception of talker gender. The importance of glottal waveform, however, was dependent on vowel quality. In the case of /u/, the glottal source nearly overpowered the effect of the k-factor. However, in the case of /ɑ/, the glottal source had very little effect; the k-factor was the major determiner of talker gender. In the case of the /i/ vowel, both glottal waveform and k-factor systematically controlled the perception of talker gender.

The present results demonstrate that glottal waveform is an important attribute of the speech signal that listeners use to identify talker gender. However, this attribute should be studied in combination with other cues since its relationship with talker gender identity did not consistently override all other cues in all conditions. It should be remembered, however, that only two glottal waveforms were used in this experiment, one male and one female. Therefore, no claims can be made to account for which aspects of the glottal waveform contributed to the perception of talker gender identification. The present study demonstrates that listeners are able to reliably identify talker gender in a manner consistent with the gender associated with the source of the glottal waveform.

The demonstration that listeners can identify talker gender on the basis of glottal waveforms formed the basis for the next experiment in which we examined the performance of listeners in identifying individual talkers from glottal information alone.

Summary and Conclusions

In the present experiment, a set of continua were created each of which ranged from male to female in six equal steps according to Fant's k-factor. All other things being equal, subjects will cross over from labeling stimuli "male" to "female" near the center of the continua. However, when the stimuli were synthesized with a male glottal waveform, listeners responded "male" further into the continua and when they were synthesized with female glottal waveforms, listeners responded "female" earlier in the continuum. These findings showed that if a sensitive

procedure is used, glottal waveform can be demonstrated to influence listener's judgements of talker gender. The ability of the glottal waveform to influence the perception of talker gender in somewhat less constrained situations will be examined in a later experiment.

Chapter 4

Talker Identification as a Function of Glottal Waveform

The data obtained in the previous experiment demonstrate that the glottal waveform is important in determining relatively gross information about talker identity -- namely, a talker's gender. Does the glottal waveform contain enough information to identify particular talkers? Although speech analysis studies demonstrate that the shape of the glottal waveform varies reliably across different talkers, the few perceptual experiments that have been conducted showed little effect of glottal waveform information on talker identification.

One set of studies indicated that when the glottal waveform from one talker was combined with the super laryngeal transfer function from another talker, listeners identified the stimulus as having been produced by the talker that contributed the superlaryngeal rather than the glottal information (Miller, 1964). In Miller's first experiment, the word "hod" was obtained from two talkers and both of the hybrid stimuli were reported to have sounded as if they were produced by the talker contributing the vocal tract transfer function. A second experiment was then conducted which coupled artificial source waveforms with a particular talker's vocal-tract transfer function using the word "hod." These source waveforms included triangle, pulse, and sinusoidal types in addition to some more realistic glottal sources. Hecker (1971) reported that listeners identified the words as coming from the producer of the vocal-tract transfer function even though listeners reported large differences in quality between the stimuli produced with different glottal waveforms. In Miller's third experiment, six speakers produced the isolated vowel /ɑ/. According to Hecker (1971), when the vocal-tract transfer functions from these utterances were combined with two artificial but realistic glottal waveforms, the perceptual differences due to the different vocal-tract transfer functions were found to be much greater than those due to different glottal waveforms. An ABX discrimination task was used in the fourth and final experiment. In this task, two natural and two hybrid samples were constructed from natural vocal-tract transfer functions and "realistic" glottal waveforms. The reference items were always two natural tokens and the test item was either a natural or a hybrid item. Listeners generally matched the hybrid items with the reference item sharing the same vocal-tract transfer functions.

These findings suggest that the glottal waveform appears to play only a very small role in talker identification. Unfortunately, it is difficult to interpret these results since the original Miller (1964) paper was an abstract. Incomplete information was provided on stimulus construction and no information was reported on the statistical reliability of these results. Most of the details of this experiment were summarized in Hecker's review (1971). Although Miller's results suggested that glottal waveform was a relatively unimportant component of talker identity, his experiments did not systematically address the relative importance of the glottal source and the vocal-tract transfer function. After summarizing Miller's studies, Hecker noted that, "the relative importance of various descriptors of the speech has not been systematically examined. Further studies along these lines could contribute to a better understanding of the acoustical manifestations of speaker identity."

In another perceptual study, Coleman (1973) demonstrated that talker discrimination can proceed relatively smoothly in the complete absence of glottal source information. The stimuli in this experiment were based on 10 male and 10 female talkers. Each of them read a 53 word prose passage using an electro-mechanical larynx. This device provided a standard 85 Hz voicing source that was common to each speaker. From these passages 5 second segments of connected speech were extracted and then placed on a test tape. The experimental tape was made up of 40 pairs of these segments in which there were 20 same-voice pairs, 10 male-female pairs, 5 male-male pairs, and 5 female-female pairs. Listeners were then presented with these stimulus pairs in random order and asked to respond whether the utterances were spoken by the same or different talkers. The results showed a very high level of discrimination. Most of the errors were contributed by only a few of the talker pairs. Confidence ratings were also recorded, and again, the low ratings were concentrated among a few talkers. Tokens from two of the female talkers were particularly difficult to discriminate. These talkers were identified as the same person nearly 40 percent of the time. Coleman also compared the identifiability of male versus female talkers and found that male talkers were easier to discriminate than female talkers.

The high performance displayed by listeners in Coleman's experiment showed that talkers may be discriminated without any glottal waveform information, suggesting that glottal waveform may be of little importance in talker identification. His methodology, however, was designed to emphasize the contribution of formant spacing. It is obvious that glottal waveform should be studied directly using more than only one glottal waveform.

The findings reported by Miller and Coleman indicate that glottal waveform appears to play a minor role in talker identification. However, other research has shown reliable differences in speech production due to glottal waveform (Carr & Trill, 1964; Monsen & Engebretson, 1977). Reliable perceptual effects of glottal waveform have also been reported for human listeners (Lass, Hughes, Bowyer, Waters, & Bourne, 1976; LaRiviere, 1975). And, performance of automatic speaker recognition systems have benefited from glottal waveform information (Wolf, 1972). Taken together, these findings, along with the earlier experiments on formant spacing, provided the impetus for the present experiment.

Since earlier research was not sensitive to the contribution of glottal waveform to talker identification, the present experiment was specifically designed to investigate these effects. Stimuli were created from natural utterances in which the effects of the superlaryngeal tract were removed by use of a pseudo-infinite length tube (Sondhi, 1975). The results of this process left the glottal waveform and fundamental frequency of the original utterances intact, while removing the formant structure -- just the converse of the stimuli used in Coleman's fixed glottal source experiment. With these stimuli in hand, we could investigate the contribution of glottal source to talker identification without formant spacing information. The testing procedure was designed so that the fundamental frequency cues were also minimized, thus leaving glottal waveform as the sole cue for talker identification.

EXPERIMENT 2

The purpose of the present experiment was to determine whether listeners use glottal waveform information to identify individual talkers. To accomplish this goal, listeners were first trained to identify a group of talkers by voice. Listeners then identified these same talkers on the basis of signals that had vocal tract resonance information experimentally removed.

Method

Subjects

Twenty-six subjects were chosen from the Speech Research Laboratory's paid subject pool for this experiment. They were paid $3.50 for a single session that lasted about one hour. None of the subjects reported any history of a speech or hearing disorder. Several had participated as paid subjects in other experiments in the Laboratory. However, the subjects were selected so that they were not involved in any work connected with the present research. All subjects were native speakers of English.

Stimuli

Training. A set of natural utterances was used for the training phase of the experiment. The ten words in this set were selected from phonetically balanced (PB) list 1 (Egan, 1948). Examples of these are given in Table 4.1. These words were spoken by six different talkers, three males and three females, who will be referred to as P, M, T, L, N, and J. The stimuli were recorded on an Ampex AG-500 audio tape recorder. The talker was stationed inside an IAC single-subject isolated acoustic chamber (Model 401-A) and read a randomized list of items into an an Electro Voice (Model EV D054 dynamic microphone positioned approximately 30 cm in front of the lips. The audio tape was then low-pass filtered at 4.8 kHz and digitized with a 12 bit A/D converter using a 10 KHz sampling rate. The input level to the A/D converter was set as high as possible without significant peak clipping over the entire list of words. Since this level was set only once per list, the individual words retained their natural relative amplitudes. When presented to subjects, the output level was set to 80 dB SPL using word 0 from talker P for calibration with a true RMS voltmeter.

Table 4.1

Natural Word List

Word number	PB number	Word
0	03	dish
1	09	bar
2	11	fuss
3	24	are
4	28	rub
5	30	deed
6	36	use
7	39	pile
8	40	rat
9	47	toe

Testing. The experimental stimuli used in the testing phase of the experiment were generated by the same six talkers as in the training phase and were intended to be exact replicas of the individuals' glottal waveforms. The glottal waveforms were extracted from the talkers with the aid of a reflectionless tube (Sondhi, 1975; Monsen & Engebretson, 1977). The resulting stimuli sound something like humming. The construction, capabilities, and limitations of the tube, as well as the rationale for selecting this method were discussed earlier in Chapter 2.

At each recording session, the talker was first allowed to practice phonating into the reflectionless tube while watching his or her glottal waveform displayed in real-time on an oscilloscope. Subjects were initially instructed to produce an extended neutral vowel into the tube. The experimenter also observed this procedure and gave hints regarding placement of the tube, strength of vocal output, and the meaning of the visual feedback. For example, if F1 was apparent in the waveform due to improper coupling, the experimenter would point this out to the talker and suggest how to cure the problem. After about 1 minute of practice, the experimenter prompted the talker to generate three stimuli: high, medium, and low in pitch. No particular frequencies were used as targets; talkers were simply asked to try to reach the upper and lower limits of their range. Each talker participated in three sessions which resulted in a total of nine different tokens of the talker's glottal waveform. The recording sessions were separated in time by at least 24 hours. Only the tokens recorded during the second session were used in the present experiment. Each utterance was amplified by an HP 465A amplifier and then digitized directly without being passed through a low-pass filter since there was no energy present above 5 KHz.

Each vocalization was recorded for 1.6 seconds and a segment was excised from this waveform that was as near as possible to 500 msec in duration. The duration was not exactly 500 msec because it was necessary for the waveform to begin and end at a zero crossing in order to eliminate onset and offset clicks. The experimental stimuli consisted of segments of the 18 utterances (three from each talker) recorded at the second session. Oscillograms of these waveforms and their Fourier power spectra are displayed in Appendices 1 and 2, respectively.

Apparatus

The present experiment was conducted in real-time with a PDP-11/34 computer system. All stimuli were output at a rate of 10 KHz and were filtered with a very steep low-pass filter at 4.8 KHz (see Klatt, 1980). Stimuli were presented

to subjects over matched and calibrated TDH-39 headphones. The trial sequence and stimulus presentation orders were determined by experiment control programs that were written specifically for each experiment.

The experimental sessions were conducted in a subject testing room equipped with six booths that allowed the simultaneous testing of six subjects. A seven button response box was stationed at each booth. In addition to the buttons, numbered 1 to 7, seven lamps were provided for feedback and one lamp was provided as a warning signal. In the conditions where the feedback lamps were used, they were illuminated over the response button that the subject should have pressed for a correct response on any given trial. Since six talkers were used in the present experiment, only six of the seven buttons and lamps were used. The experimenter was present during testing and controlled the computer from a separate booth within the subject testing room.

Procedure

Each experimental session was subdivided into three phases, familiarization, training, and testing. The first two phases were used to train the listeners to identify the natural voices of the six talkers. The last phase was conducted to assess the identification of the glottal waveforms of these six talkers. This methodology has been used successfully in training new linguistic contrasts (see McClasky, Pisoni, & Carrell, 1983; Pisoni, Aslin, Perey, & Hennessey, 1982).

Familiarization. During the familiarization phase, the listeners were presented with a natural token of each voice in the following sequence. First, a warning light would be illuminated on the response box for 500 msec to indicate the beginning of a new trial. After a delay of another 500 msec, an utterance from the natural word list was presented over each subject's headphones. Following a final 500 msec delay, a lamp was illuminated for 250 msec over button 1 on the response box. This indicated that the talker just pres ented should be associated with button 1. After a 3 second delay, the warning lamp went on to signal the beginning of the next trial. Next, the same word was presented again, however, this time spoken by talker 2. The lamp over button 2 was then illuminated. This sequence was repeated for all six talkers before moving on to the next word. Ten words for a total of 60 trials were presented in this fashion.

During the entire familiarization procedure no overt response was required from the listeners. They simply listened to the voices and watched the feedback lights as

the stimuli were presented in a fixed sequence. The familiarization phase lasted about 5 minutes.

Training. The training phase began after a short review of the instructions by the experimenter. The same stimuli used during familiarization were presented here, although, in this phase, they were presented in a random order. On each trial the listener received a warning signal followed by a 500 msec delay. A test word was then presented over the headphones. Listeners had up to 4 seconds to respond with a button press. After all the listeners had responded (or after 4 seconds had elapsed), the feedback lamp over the correct response was illuminated for 750 msec which completed the trial.

Four repetitions of each stimulus word were presented for each talker during the training phase for a total of 240 trials. This phase of the experiment lasted approximately 20 minutes.

Testing. After a short break, the testing phase of the experiment began. The trial sequence was identical to the procedure used in the training phase with the exception that glottal waveform stimuli were used and no feedback was provided after listeners entered their responses. Listeners were presented with one stimulus per trial and were required to press the button corresponding to the the talker who produced the utterance. Listeners were told that most of the sounds that they would be hearing would sound something like "humming" and that they were produced by the same six talkers that they had just learned. Listeners were also told that this task was much more difficult than the training task because each talker would make these humming sounds at a number of different pitches. The listeners were instructed to respond on each trial even if they had to guess.

In this phase of the experiment, 24 unique stimuli were presented three times each in two blocks for a total of 144 trials. The blocks were separated by a 1 minute break. Each of the six talkers contributed four items to the list. The first three items were samples of the low, medium, and high pitched glottal waveforms of each talker. Three pitches were used so that talker identification could not be based exclusively on fundamental frequency. However, only the medium frequency stimuli were actually scored. The fourth item was a word from the natural word list spoken by the same talker. These items were included at random intervals as "probe trials" throughout the testing session to remind listeners of the natural speech quality of the talkers that they were required to identify. These stimuli also served as a useful control to compare identifications of glottal waveform with.

Results

Training

It was surprisingly difficult to train listeners to identify the natural voices of the six talkers in a one hour experimental session. Based on post-test interviews, listeners generally used two steps in this task. First, they determined whether the voice was male or female, and then they determined the specific identity of the individual talker. Listeners reported no trouble distinguishing male from female stimuli. However, the selection of the specific talker within each category was more difficult. As noted earlier in the procedure section, the training method that was used in this experiment was one that had been successfully used in earlier speech perception experiments. Although listeners had many exposures to the voices, they did not learn to identify the natural intact utterances of the six talkers at asymptotic levels in the time allotted. Figure 4.1 illustrates the mean accuracy level for each talker for this phase of the experiment.

All listeners performed well above chance (17%) on the identification of each talker. The worst performance (55%) was observed with talker J [$t(25) = 13.24$, $p < .0001$]. Even if chance is defined as 33% because of the assumption that subjects would be perfect at segregating male from female talkers, this result was significantly above chance [worst case $t(25) = 7.67$, $p < .0001$]. Combined over all conditions the accuracy level was 79%.

As Figure 4.1 clearly shows, the identification performance was not evenly distributed across the six talkers. The male talkers were identified more accurately than the female talkers [$F(1,25) = 112$, $p < .0001$]. An interaction between gender and talker [$F(2,50) = 12.08$, $p < .0001$] indicated that even within a gender class the difficulty of talker identification varied.

The confusion matrix in Table 4.2 presents a more detailed picture of the pattern of listener responses. First, it can be seen that most of the incorrect responses remained within the appropriate gender class. Second, the female talkers were more confusable with each other than the male talkers, especially talkers L and J.

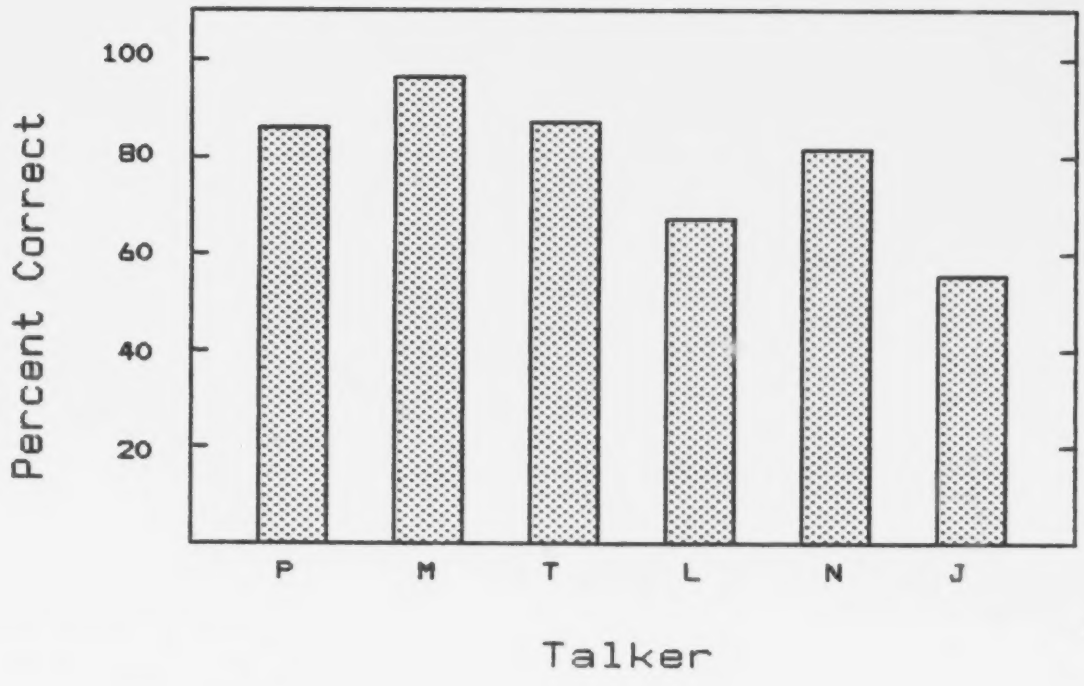


Figure 4.1. Natural voice identification accuracy during the training phase for each of the six voices. The male voices are shown in the left group and the female voices are shown in the right.

Table 4.2

Natural Word Confusion Matrix

Stimulus	Response						
	P	M	T	L	N	J	None
P	788	34	95	0	1	1	41
M	20	888	22	0	0	0	30
T	70	48	823	2	1	0	16
L	0	0	2	606	94	216	42
N	0	0	0	96	746	81	37
J	0	0	0	350	70	501	39

Note. The maximum number of responses per stimulus was 960.

Identification

Despite the apparent difficulty of the training task, listeners' performance was high enough to warrant examining the results of the glottal identification testing phase of this experiment. Identification performance on this task is shown in Figure 4.2. These scores show only the listeners' performance on the medium pitch glottal waveforms. High and low frequency waveforms were included as filler items to prevent listeners from using fundamental frequency as the only cue to talker identification. Since six response alternatives were available, chance was assumed to be 17%.

Upon examination of Figure 4.2, it is clear that substantial differences in response accuracy are present among different talkers. Listeners were reliably above chance in identifying the glottal waveforms extracted from talkers P, M, T, and L [$t(25) = 2.56$, $p < .02$; 6.23, $p < .0001$; 5.70, $p < .0001$; and 4.07, $p < .0005$ respectively]. However, identification was below chance for talkers N and J [$t(25) = .19$, n.s.; and $t(25) = .54$, n.s.].

The performance of listeners on the natural utterances that were interspersed along with the glottal waveform stimuli throughout the testing phase is shown in Figure 4.3. This figure illustrates the performance of the subjects on talker identification based on intact utterances. The results were obtained with the same stimuli that were presented during the training phase and are in agreement with the testing phase data. A moderate correlation was found between the training and testing performance across all talkers [$r = .54$, $t(25) = 3.14$, $p < .005$]. The training data provided a more sensitive indication of the relative difficulty of the individual talkers since it was also an indirect measure of learning time. The testing phase results showed that, first, the subjects had learned to identify the male talkers by the beginning of the testing phase (in fact, the mean accuracy was the same for each male voice -- exactly .91), and second, that while the female talkers were identified at levels well above chance, performance in identification of their voices was significantly poorer than performance with the male voices [$F(1,25) = 49.5$, $p < .0001$].

The confusion matrices shown in Tables 4.3 and 4.4 further illustrate the response patterns for the glottal waveforms and the natural utterances, respectively. Note the clear pattern of results in the natural utterance identification data shown in Table 4.3. Identification performance for the male talkers (P, M, and T) was excellent

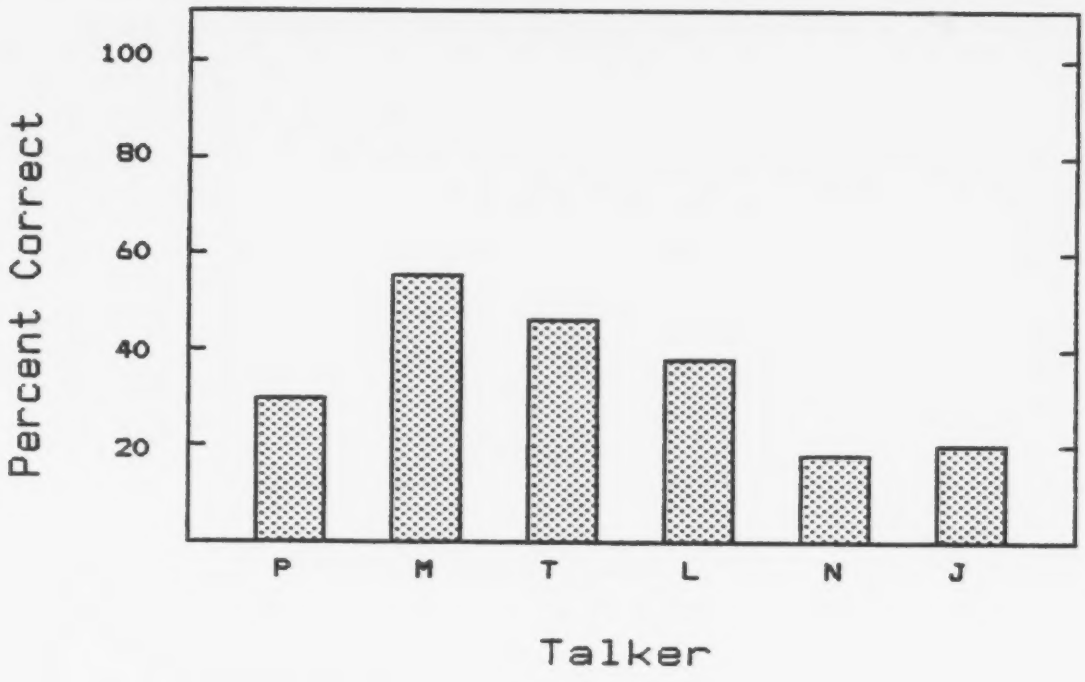


Figure 4.2. Glottal waveform identification accuracy during the testing phase for each of the six talkers.

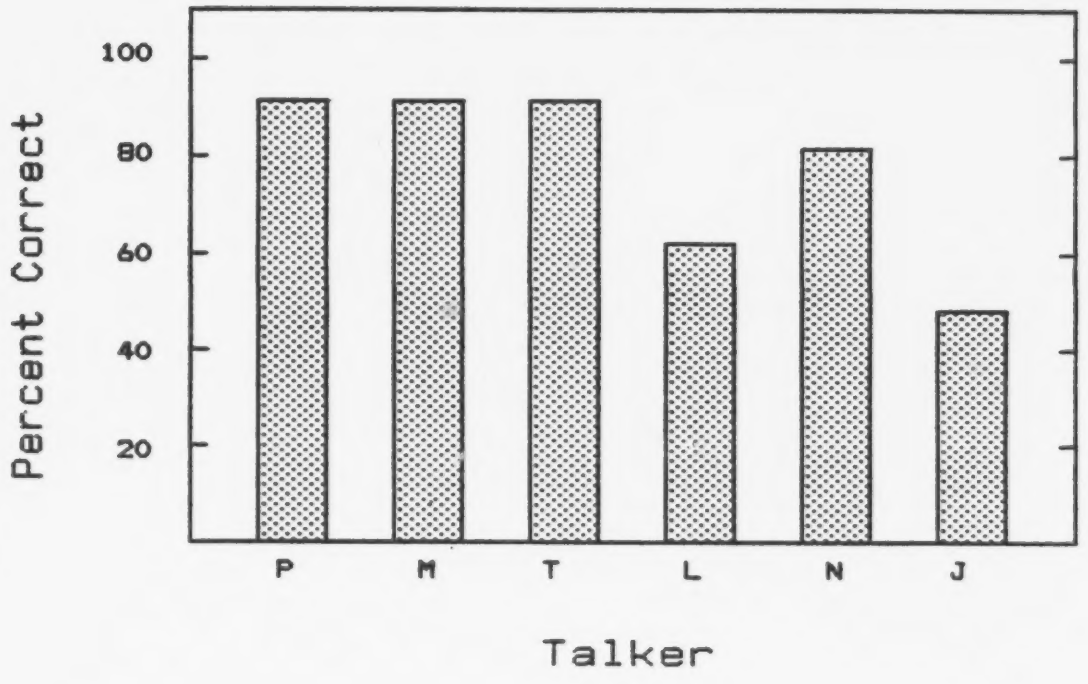


Figure 4.3. Natural voice identification accuracy during the testing phase for each of the six voices.

Table 4.3

Natural Word Identification Confusion Matrix

Testing Phase

	Responses						
Stimulus	P	M	T	L	N	J	None
P	122	2	10	0	0	0	6
M	8	123	4	0	0	0	5
T	4	6	124	0	0	0	6
L	0	0	0	78	16	32	14
N	0	1	0	13	111	10	5
J	0	0	0	66	2	59	13

Note. The maximum number of responses per stimulus was 960.

Table 4.4

Glottal Waveform Identification Confusion Matrix

	Responses						
Stimulus	P	M	T	L	N	J	None
P	40	20	44	14	4	7	11
M	34	72	20	1	1	1	1
T	40	15	62	4	3	4	12
L	16	1	19	47	13	26	18
N	9	3	12	45	23	26	22
J	12	7	36	22	18	25	20

Note. The maximum number of responses per stimulus was 960.

as shown by the large numbers on the diagonal, no systematic pattern was evident in the distribution of errors.

Identification performance for the female talkers was, however, substantially worse. The error patterns revealed that talkers L and J were especially confusable with each other. This table also demonstrates that virtually no confusions occurred in identification between male and female talkers. Table 4.4 also shows a similar although much weaker pattern for the identification of the glottal waveforms. Again, few confusions were observed in errors between male and female talkers. Although poor overall, identification of male talkers was still much better than identification of female talkers.

Discussion

Taken together, the results from the present experiment demonstrate that some cues to talker identification are preserved in the glottal waveform and that, for voices that were well learned, this information is sufficient for the identification of talkers by human listeners at levels well above chance. Listeners in this study were able to identify talkers on the basis of their glottal waveform at levels above chance for all but the two female talkers who were well learned during training. These findings are in sharp contrast with the earlier results of Miller (1964) and Coleman (1973) who found that glottal waveform provides minimal cues for talker identification. The major differences in the results appears to be due to the fact that these earlier experiments were biased towards showing the importance of formant spacing, whereas the present experiment was oriented toward demonstrating the importance of glottal waveform in talker identification. None of these experiments, including the present one, was designed to assess the relative importance of or the potential interactions between these cues to talker identification. The last experiment in the present investigation was specifically designed to address this issue.

The fact that two talkers could not be identified on the basis of their glottal waveforms was attributed to the poor learning of those talkers by voice. However, it is also possible that the glottal waveforms of those talkers were simply non-discriminable. In order to determine whether, in principle, it would be possible to identify talkers based on their glottal waveform characteristics alone, a discriminant analysis was performed on the glottal sources. First, a Fourier transform was performed on each of the medium-frequency glottal waveforms that had been collected. These included the tokens from all three days.

The amplitudes of the first 16 harmonics of the energy spectrum (measured in decibels) from this transform served as the input to the analysis. Correct classification of each of the spectra to the appropriate talker was found in 100 percent of the cases. That is, the discriminant analysis procedure was able to classify glottal waveform to the appropriate talker perfectly, indicating that the energy spectrum of the glottal waveform contained enough information to correctly identify talkers. However, one should bear in mind that the discriminant analysis was performing a much different, and in some ways simpler, task than the listeners in the perceptual experiment. The discriminant analysis was both "trained" and "tested" on glottal sources, whereas the listeners were trained on naturally spoken utterances and then tested on the glottal sources. In any case, this analysis demonstrates that the information in the spectrum of the glottal source is sufficient to specify the talker. Thus, in principle, enough talker specific information is present in the glottal waveform to allow accurate talker identification.

Another finding that emerged from the present experiment was that 20 minutes was simply not enough time to learn to identify six talkers by voice on the basis of single word utterances. This finding, no doubt, depends on the discriminability of the specific voices to be learned. In the present case, we found that female voices were more difficult to learn to identify than male voices. Obviously, with only three male and three female talkers, it is difficult to draw any firm conclusions about the relative identifiability of male and female talkers. In his experiment on the identification of talkers from only formant information, Coleman (1973) found that female talkers were, in fact, also more difficult to identify than males. Further investigation of these differences is clearly warranted.

Although the present results did show significantly poorer glottal waveform identifiability for female talkers than for male talkers, they cannot be used to decide whether female glottal waveforms would have been just as discriminable if the female speakers had been learned to an equivalent degree during training on natural words. More training time will be necessary for each talker and more talkers will be required in order to answer this question.

The findings obtained in this experiment indicate that glottal waveform is an important component of the perception of talker identity, but they do not show the relationship between glottal waveform and any other cues to talker identity. The next two experiments were designed to investigate how glottal waveform interacts with fundamental frequency and formant spacing.

Summary and Conclusions

In the present experiment, subjects were trained to identify six talkers by voice. They were then tested on their ability to identify these talkers on the basis of glottal waveforms independent of superlaryngeal filtering. The results showed that they were capable of doing this for those talkers who were well learned. Thus we conclude that glottal waveform information is sufficient for talker identification at levels above chance.

Chapter 5

Naturalness and Intelligibility of Synthetic Talkers

Since synthetic stimuli were required for the final experiment of the present investigation, it was necessary to know whether the synthesis techniques currently in use are capable of modeling an individual's voice to a sufficient degree to make the modeling of specific cues a meaningful activity. Although the quality of synthetic speech has improved substantially over the past decade, the perceptual testing that has been conducted has been primarily concerned with the segmental intelligibility of the message presented to the listener. Research is still needed on the factors that influence the perceived naturalness of synthetic speech. In the case of the talker identification, an obvious experiment would be to measure listeners' accuracy in a talker identification task using appropriately constructed stimuli. Such an experiment was conducted and will be presented in Chapter 6. Before this experiment was carried out, however, it was necessary to perform a perceptual experiment of a more general nature in order to examine the acceptability of the synthetic stimuli. The present experiment was designed to asses listeners' impressions of the naturalness and intelligibility of synthetic speech produced by a modified version of the Klatt software synthesizer using a specific set of stimulus construction methods.

The synthetic stimuli used in the next two experiments were generated by the modified Klatt synthesizer (see Chapter 2). While it has been claimed that this synthesizer is capable of creating an utterance that is "virtually indistinguishable from the original in both intelligibility and naturalness" (Klatt, 1980, p. 985), this claim has not been validated with perceptual testing using human listeners. Furthermore, in order to achieve excellent results, it is necessary to use time consuming analysis-by-synthesis methods rather than one-step parameter calculation processes. That is, it is not possible to simply measure certain speech parameters such as formant frequency and bandwidth from an utterance and then enter these values into the synthesizer if one hopes to mimic both the intelligibility and the naturalness of the original talker. Unfortunately, this makes it difficult to automatically generate high quality intelligible and natural sounding speech with well defined and well understood parameters (Klatt, 1980). Furthermore, if excessive hand manipulation of the parameters is necessary, then the utility and generality of the parameters being manipulated is weakened

substantially. Because of these considerations, it is necessary to question whether any experiments that use such a tool to study talker identity are capable of reaching meaningful conclusions even when reasonable care and several iterations of analysis and synthesis are used.

One way to approach the problem of perceived natural ness of synthetic speech is simply to ask listeners to rate the naturalness and intelligibility of the speech on an arbitrary scale. While such an experiment may not be especially sensitive -- due to the subjects' reactions to the task demands -- this methodology may provide us with some insights into the general question posed.

The synthetic speech used in this experiment was generated in the following manner. A natural utterance was first analyzed both spectrographically and by linear predictive coding (LPC) methods and the relevant synthesis parameters were extracted. These parameter values were then used as input to the modified Klatt synthesizer which performed the actual synthesis of the speech waveform. The resulting speech was then analyzed both by the experimenter, for phonetic quality, and by LPC methods, for match to the original utterance. Depending on the utterance and talker being modeled, this process was repeated a varying number of times. The goal of this synthesis strategy was not to provide a perfect spectral match in an LPC sense which would certainly preserve intelligibility and naturalness as would an analog tape recording but rather to model the formant structure, fundamental frequency, and glottal source as closely as possible since these were the synthesis parameters of interest in our investigation of the cues to talker identification.

The mean ratings of naturalness and intelligibility obtained from the listeners may be used as an indicator of the overall quality of the synthesis procedure. These perceptual data should therefore reflect the validity and sufficiency of the parameters that were extracted and manipulated to represent the cues that were presumed to control naturalness and intelligibility. In addition, these ratings should provide more detailed information on the differences in perception between the particular talkers that were modelled by these techniques.

The results from these analysis-by-synthesis procedures allow the examination of some additional interesting questions. First, are subjects' judgements of intelligibility and naturalness correlated? That is, are the specific utterances and talkers that are better synthesized for intelligibility also better synthesized for perceived naturalness? Second, are intelligibility and naturalness equally well modelled by the present methods? And, third, are there differences in the synthesizer's

ability to model different talkers, both in terms of naturalness and intelligibility? The present experiment was carried out to address these issues.

EXPERIMENT 3

Method

Subjects. Twenty subjects were chosen from the Speech Research Laboratory's paid subject pool for this experiment. They were paid $3.50 for a one-hour session. The data from six of the subjects was discarded because they failed to completely fill out their response forms as required. All analyses were conducted on the data from the remaining 14 subjects. None of the subjects reported any history of a speech or hearing disorder. Several had participated as paid subjects in other research in the Laboratory although they were not involved in any studies connected with the present experiment. All subjects were native speakers of English.

Stimuli. Two sets of stimuli were used in this experiment. The first set was taken from the natural speech database used for the training phase of Experiment 2. This set consisted of 10 words (listed in Table 4.1) spoken by three male and three female talkers, referred to as P, M, T, L, N, and J. As described earlier in the training section of Experiment 2, the stimuli were first recorded on audio tape using an Electro Voice EV D054 dynamic microphone and a professional quality Ampex tape recorder. This tape was then digitized at a 10 KHz sampling rate and stored on a PDP11/34 computer.

The second set of stimuli were synthetic versions of these 60 natural utterances. These words were synthesized using the modified version of the Klatt software synthesizer so that they retained the fundamental frequency, formant spacing and glottal waveform characteristics of the natural voice that they had been derived from.

As mentioned earlier, the synthetic stimuli in this experiment were produced using an iterative analysis-by-synthesis procedure. The natural utterance was first analyzed to extract the three cues under investigation. In addition to formant spacing, fundamental frequency, and glottal waveform, the durations, formant bandwidths, and formant amplitudes were also extracted and retained for each model word produced by each talker.

Two audio tapes consisting of 300 stimulus pairs were generated with an audio tape making program designed specifically for this purpose. The stimuli were output from the computer at a sampling rate of 10 kHz through a 4.8 kHz low-pass filter. The first item in each stimulus pair was a digitized natural token of a word, the second item was the synthetic version of the same word. The interstimulus interval was one second and the intertrial interval was seven seconds. A different random ordering of the 300 stimulus pairs was used on each of the two test tapes.

Apparatus. The audio tapes were presented to subjects via an Ampex AG-500 tape recorder identical to the one used to record the natural stimuli. The tape recorder output was amplified and presented over TDH-39 headphones at an average signal level of 80 dB SPL. The subject's responses were recorded on prepared answer sheets similar to the one shown in Figure 5.1. These responses were then transferred to computer storage after the experimental sessions for further analysis. The experimental sessions were conducted in two testing rooms simultaneously. Each room contained six subject booths each of which was equipped with with a small desktop and pair of headphones.

Procedure. The trial sequence was determined randomly by the audio tape making program described above. Listeners were seated in both experimental testing rooms and were given written instructions that described the experiment and procedure.

Since the task used in this experiment involved subjective judgements about both intelligibility and naturalness, the specific details of the instructions might influence subjects' responses. Because of this, the text of their instructions is reproduced below:

> Welcome to the Speech Research Laboratory. We have an interesting experiment planned for today. You will hear both natural speech (produced by humans) and synthetic speech (produced by a computer). We are trying to improve synthetic speech and make it sound like particular individuals. To do this, we have developed a large number of stimuli and we would like you to rate them for us.
>
> This experiment consists of 300 trials. On each trial you will hear one word, spoken twice. The first occurrence will be natural and the second will be synthetic. We would like you to rate how the computer sounds in comparison to the human. (That is, how the second word sounds in comparison to the first.)
>
> You will notice that on the answer sheets there are two responses required for each trial;

Name_____

Trial #	Unintelligible				Intelligible			Unnatural				Natural		
1	1	2	3	4	5	6	7	1	2	3	4	5	6	7
2	1	2	3	4	5	6	7	1	2	3	4	5	6	7
3	1	2	3	4	5	6	7	1	2	3	4	5	6	7
4	1	2	3	4	5	6	7	1	2	3	4	5	6	7
5	1	2	3	4	5	6	7	1	2	3	4	5	6	7
6	1	2	3	4	5	6	7	1	2	3	4	5	6	7
7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
8	1	2	3	4	5	6	7	1	2	3	4	5	6	7
9	1	2	3	4	5	6	7	1	2	3	4	5	6	7
10	1	2	3	4	5	6	7	1	2	3	4	5	6	7
11	1	2	3	4	5	6	7	1	2	3	4	5	6	7
12	1	2	3	4	5	6	7	1	2	3	4	5	6	7
13	1	2	3	4	5	6	7	1	2	3	4	5	6	7
14	1	2	3	4	5	6	7	1	2	3	4	5	6	7
15	1	2	3	4	5	6	7	1	2	3	4	5	6	7

Figure 5.1. Response sheet for collecting subject's ratings of naturalness and intelligibility in Experiment 3.

intelligibility and naturalness. In order to rate intelligibility, circle 7 if the second word is just as intelligible as the first. The word does not have to sound human, just intelligible. Circle number 1 if you cannot understand the second word at all. Use the numbers between 1 and 7 to rate the intelligibility between these two extremes. The second response required is a rating of naturalness. Circle number 7 if the second word sounds like it was produced by the same talker as the first word. Circle number 1 if it sounds like it was produced by a very different talker (for example a machine or animal).

If you have any questions please feel free to ask the experimenter. Thank you for your participation.

These instructions were also read aloud by the experimenter immediately before the trial sequence. After answering any questions, the experiment was initiated.

Results and Discussion

The results of the present experiment indicated that listeners found the synthetic speech highly intelligible and natural. In support of this statement, the naturalness and intelligibility ratings collected for each talker are shown in Figure 5.2. Each bar is a mean rating averaged across all listeners. Recall that listeners were required to select a number from 1 to 7 to rate the stimuli on both measures. The solid bars represent intelligibility scores and the speckled bars represent naturalness scores. An examination of this figure reveals that these ratings (grouped by talker) all lie between of 3.5 and 7.0 on the response scale. Thus, listeners found the synthetic speech acceptable on both of measures. Of course, there is nothing special about a rating of 3.5 that would cause it to be a threshold of acceptability, but it does allow an overall characterization of the listeners' subjective impressions of naturalness and intelligibility of these stimuli.

Another clear pattern shown in Figure 5.2 is that the intelligibility ratings were consistently higher than the naturalness ratings for all six talkers. The effect was statistically significant as revealed by a main effect of naturalness versus intelligibility in a two-way analysis of variance [$F(1,13) = 34.85$, $p < .0001$]. In considering these results, it should be kept in mind that naturalness and intelligibility are very dissimilar measures. There is no reason to suppose that a value on one scale directly corresponds to the same value on the other. However distant the correspondence, the strength and consistency of these

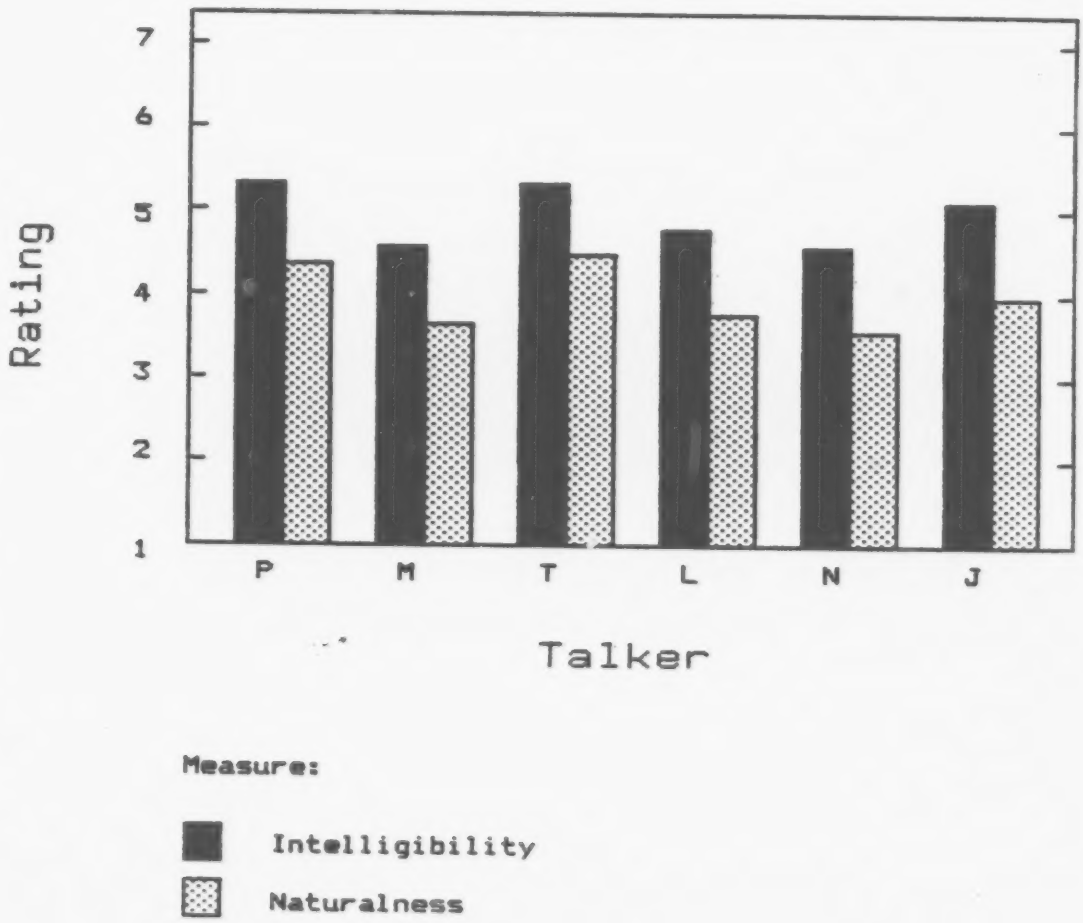


Figure 5.2. Mean ratings of intelligibility and naturalness averaged across all words. Each pair of bars shows the ratings for one synthetic voice; the left-hand bar of each pair represents intelligibility and the right-hand bar represents naturalness.

results argue that the ratings of intelligibility of the synthetic utterances were much higher than the ratings of naturalness.

The correlations between the naturalness and intelligibility ratings for the different talkers is shown in Table 5.1. These correlations are based on the identification and naturalness ratings for each talker averaged over the ten words in the stimulus list. When examining this table it should be noted that if any single correlation was compared to zero, it would have to be greater than .532 to be significant at a probability of less than .05 (two-tailed). The correlations between the intelligibility ratings of each talker with his or her naturalness rating were all relatively high, as shown by the diagonal (printed in boldface). Each one of these within-talker correlations is significantly different from zero.

The correlation matrix is especially useful in examining the detailed results of the present experiment. However, even though the correlations shown in Table 5.1 appear reasonably large, claims cannot be made regarding some of the more complex patterns in the data without conducting further analyses. Because of this, multivariate techniques were used to supply interpretations of the major results of the present experiment. The outcomes of these procedures will be used to argue that the synthesis techniques captured important talker specific qualities found in natural speech.

Table 5.2 is a summary of the results of a multiple correlation analysis conducted separately on intelligibility and naturalness scores that were grouped by talker. The scores in the upper panel refer to the intelligibility intercorrelations; the scores in the lower panel refer to the naturalness intercorrelations. For both groups, the first column contains the initial of the talker under consideration. The second column contains the squared multiple correlation (R^2) between that talker and the other five talkers. (This R^2 was a measure of the correlation of the predicted intelligibility scores with the actual intelligibility scores. The predicted scores were based on the data from the other five talkers and the actual scores were those obtained in the experiment itself.) These correlations were then transformed into F-statistics in column 4 (with 5 and 8 degrees of freedom) and finally into significance levels in column 5, indicating the fit of the regression equations.

The multiple regression analysis revealed that the intelligibility ratings of different synthetic talkers were highly intercorrelated. Any given talker's intelligibility rating could be predicted at levels significantly above chance from the intelligibility ratings of the remaining

Table 5.1

Naturalness and Intelligibility Score Correlation Matrix

	Intelligibility					
	P	M	T	L	N	J
Naturalness						
P	.722	.575	.693	.627	.597	.636
M	.567	.684	.640	.568	.435	.510
T	.739	.693	.860	.769	.655	.781
L	.474	.446	.545	.624	.447	.655
N	.614	.572	.617	.685	.650	.754
J	.357	.402	.428	.585	.410	.640

Table 5.2

Multiple Correlations of Each Talker with All Other Talkers

Talker	Squared Multiple Correlation	Multiple Correlation	F-Statistic	Significance
Intelligibility				
P	.94437	.97179	27.16	.00008
M	.84952	.92169	9.03	.00382
T	.87871	.93739	11.59	.00168
L	.98172	.99082	85.93	.00001
N	.94677	.97302	28.46	.00007
J	.97686	.98836	67.55	.00001
Naturalness				
P	.92798	.96332	20.62	.00022
M	.66244	.81390	3.14	.07322
T	.82432	.90792	7.51	.00685
L	.97487	.98736	62.08	.00001
N	.93690	.96794	23.76	.00013
J	.97582	.98784	64.58	.00001

talkers. A similar finding also held true for naturalness ratings. In the case of naturalness ratings, however, talker M deviated slightly from this general pattern. For this particular talker, the naturalness ratings of the other five talkers predicted his scores with a significance level of only .075. Despite this one marginal finding, the general pattern, grouped by talker, indicated that intelligibility scores were significantly correlated with other intelligibility scores and that naturalness scores were significantly correlated with other naturalness scores.

These findings appear to be related to two factors. First, the results may reflect the natural and systematic differences in speech production between talkers, which were well modeled by the synthesis procedure. Alternatively, the findings could be interpreted to mean that the synthesis procedures were differentially capable of modelling the six voices both in terms of naturalness and intelligibility. In the case of speech intelligibility, it is well known that talkers differ substantially in the intelligibility of their speech (Hood & Poole, 1980). Thus, the first explanation is probably sufficient to account for the variability in the observed data. Although it is possible that the differential ability of the synthesis system to model different voices could underlie these results, another experiment, examining ratings of intelligibility of the same natural words produced by the same talkers would be necessary to resolve this question definitively. Up to the present time, we have been unable to find a study reported in the literature that systematically examined differences in naturalness between talkers. Based on our observations, we would expect, however, that talkers should differ systematically on this measure as well. In any case, the present findings are consistent with the assumption that the talker differences were preserved reliably by the present synthesis techniques.

The results of the multivariate regression analyses strongly support the conclusion that the synthetic stimuli adequately model important talker-specific qualities. Since synthesis procedures using a number of components of a talkers voice quality appeared to reflect important natural talker differences, these results support the assumptions of separately defining the cues that were manipulated to model individual talker differences.

In the speech perception literature, researchers have often been concerned with the question of whether naturalness and intelligibility are independent from one another. Indeed, the earliest synthesis with the Haskins pattern playback (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952) produced highly intelligible, but very unnatural sounding synthetic speech. The pattern of correlations observed in the data between naturalness and

intelligibility ratings are rather unclear on this issue, and therefore a multivariate linear regression was performed to examine this question in greater detail. The results of this analysis are shown in Table 5.3. The table has been divided into two parts. In the top half of the table, the results of an analysis in which naturalness scores based on each talker were used to predict the intelligibility of the different talkers is shown. In the bottom half of the table, the results of the complementary analysis, in which intelligibility scores were used to predict the naturalness of the different talkers is shown. The R^2 values are shown in column 2, the F-statistics are shown in column 4 (with 6 and 7 degrees of freedom), and the significance levels are shown in column 5. An examination of these values reveals that naturalness scores cannot be used to reliably predict intelligibility scores, and, conversely, that intelligibility scores cannot be used to predict naturalness scores.

The assertion that the naturalness and intelligibility of the different synthetic talkers were two relatively independent factors is further supported by a principle component analysis conducted on the correlation matrix of the naturalness and intelligibility scores for different talkers shown in Figure 5.1. In this analysis, two underlying factors (linear combinations of the 12 input factors) accounted for 87% of the total variance. Factor 1 was most strongly correlated with the six intelligibility factors, and Factor 2 was most strongly correlated with the six naturalness factors. That is, the overall correlation matrix could be well described by the assumption of two independent factors, one corresponding to naturalness and one corresponding to intelligibility. While this particular analysis is only descriptive in nature, it does provide additional support for the proposal that naturalness and intelligibility are independent and separable components of a talker's voice.

Summary and Conclusions

The findings from the present experiment showed that listeners found the synthetic stimuli to be acceptable in terms of naturalness and intelligibility ratings. Although subjects rated naturalness well below intelligibility, the mean ratings were in the upper portion of the rating scale. In addition, we found that the two measures were relatively independent of each other suggesting that the cues underlying the perception of talker identity may be separate from those used to support segmental intelligibility of the linguistic message.

Table 5.3

Multiple Regression on Naturalness and Intelligibility

Talker	Squared Multiple Correlation	Multiple Correlation	F-Statistic	Significance
Intelligibility Dependent				
P	.73467	.85713	3.23	.07526
M	.69679	.83474	2.68	.11147
T	.86786	.93159	7.66	.00836
L	.63834	.79896	2.06	.18329
N	.63351	.79593	2.02	.19006
J	.69534	.83387	2.66	.11302
Naturalness Dependent				
P	.65946	.81207	2.26	.15517
M	.69722	.83499	2.69	.11102
T	.84995	.92193	6.61	.01265
L	.66420	.81499	2.31	.14918
N	.70857	.84177	2.84	.09933
J	.71179	.84368	2.88	.09615

In the final experiment, reported in the next chapter, the same synthesis methods were used to systematically manipulate the cues to talker identity. The relationship between these attributes and ratings of naturalness were also examined in greater detail.

Chapter 6

Talker Identification using a Factorial Combination of Cues

It is a commonplace finding in cognitive psychology that information is produced, perceived, and stored redundantly (e.g. Lashley, 1950). In the field of speech perception, for example, it is now uniformly accepted that linguistic information is not encoded in the speech stream with simple, unitary acoustic cues that unambiguously specify phonetic segments but rather that successive phonemes are encoded in the speech waveform with a great deal of overlap and redundancy (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). It seems reasonable to suppose that nonlinguistic information such as talker identity is also not likely to be encoded with single cues or attributes in the speech signal. Indeed, from an evolutionary perspective, there is every reason to expect that the indexical properties of language are marked redundantly in the signal just as the linguistic properties are. In order to study talker identification, it seems appropriate to first isolate several important cues and then to examine the effect of their mutual interactions on perception. The capabilities of modern speech synthesis systems combined with the methodology of experimental psychology are well suited to this task. When a sufficient number of separate cues have been identified and their interactions well specified it should be possible, at least in principle, to model the perception of talker identity just as researchers have done with the perception of phonetic segments which carry the bulk of linguistic information in the speech signal.

The goal of the present experiment was to examine the interaction of the three acoustic attributes we have identified for the perception of talker identity: fundamental frequency, formant pattern, and glottal waveform. Although other attributes are undoubtedly useful, these three were chosen because previous studies had shown them to be valuable in this regard and because fundamental frequency, formant spacing, and glottal waveform are all components that are present in word and syllable length utterances.

In order to examine their combined effects on the perception of talker identity, a large set of words was synthesized with a factorial design. The design included manipulations of the fundamental frequencies, formant patterns, and glottal waveforms of two male and two female talkers. These manipulations produced a stimulus set in

which most of the items contained conflicting cues, that is, simultaneous cues for more than one talker. Therefore, the identification of a particular token of the speaker depended on the relative perceptual importance of the cues.

The identification experiment used with this stimulus set was relatively straightforward. Subjects were first taught to identify, by voice, the four talkers that had been modeled in the construction of the stimuli. This was accomplished through the use of natural word tokens from each talker. Subjects were then required to identify the talker for each of the synthetic test words.

Naturalness ratings were also collected on each trial in this experiment. As shown in the previous experiment, ratings of naturalness and intelligibility are relatively independent from one another; that is, it is possible for the speech to be highly intelligible but also very unnatural. Is naturalness also independent of talker identity? Can a listener make accurate judgements of talker identity but still perceive the speech as unnatural or mechanical? Naturalness ratings provided one way to answer these questions. An analysis of these ratings also provided a way to assess the relationship between naturalness and the three talker-specific attributes that we manipulated in this study.

EXPERIMENT 4

Method

Subjects

Thirteen subjects were selected randomly from the Speech Research Laboratory's paid subject pool for this experiment. The subjects were paid $5.00 for a two-hour session on Day 1, $5.00 for a two-hour session on Day 2, plus a $4.00 bonus for participating in both sessions. None of the subjects reported any history of a speech or hearing disorder. Several had participated as paid subjects in other experiments in the Laboratory but they had not been involved in any work connected with the present research. All subjects were native speakers of English.

Stimuli

Three sets of stimuli were used in this experiment: natural, synthetic, and factorial combination stimuli. The natural stimuli were a subset of the stimuli that were used in the training phase of Experiment 2. Briefly, this stimulus set consisted of ten monosyllabic words taken from PB list 1. They were spoken by four of the six talkers from Experiment 2 (P, T, L, and N). Four talkers were chosen for this experiment rather than six due to the difficulty subjects had in the training task in Experiment 2. Talker J was removed due to her confusability with talker L, and talker M was removed so that there would be an equal number of male and female talkers. Talker M was also the most discriminable of the male talkers and the removal of his voice helped reduce some of the differences in the ease of learning between the male and female natural voices.

The second set of stimuli was a subset of the stimuli used in Experiment 3. These consisted of synthetic versions of stimuli 3 through 9 from the list of the ten PB words shown in Table 4.1. Tokens of the seven words synthesized with the characteristics of four different talkers produced a total of 28 unique stimuli. These stimuli were synthesized with the original talker's fundamental frequency, pitch contour, formant pattern, and glottal waveform.

The third set of stimuli, the factorially combined cue set, were specifically designed for this experiment. This set consisted of a total of 240 different items. Three different words: dish, bar, and fuss were synthesized with all possible combinations of the four talkers' formant patterns and glottal waveforms. Each of these was synthesized at five different fundamental frequencies. The frequencies spanned the range used by the four natural talkers in five equal steps. Additionally, a 15% linear drop in F0 across the duration of the word was added in order to hold the fundamental frequency contour information constant across all talkers while preserving a certain degree of naturalness. The frequencies chosen were: 110-95 Hz, 140-120 Hz, 170-145 Hz, 200-170 Hz, and 230-195 Hz. These 240 stimuli were synthesized on the modified version of the Klatt software synthesizer described in Chapter 2.

Apparatus

A PDP 11/34 computer was used to present stimuli, collect responses, and control all experimental events. All stimuli were output at a rate of 10 KHz, low-pass filtered at 4.8 KHz, and then presented to subjects at an average of 80 dB SPL over matched and calibrated TDH-39 headphones.

The experimental sessions were conducted in the the same manner as in Experiment 2. One additional piece of equipment, a 12 inch black and white CRT video display was added in the testing phase of this experiment. The video monitor was driven by a VIURAM model V11, 80 by 24 character generator and was used to display alphanumeric information to each subject.

Procedure -- Day 1

The experiment was conducted in two two-hour sessions on consecutive days. On the first day, the session was divided into four phases: familiarization, natural-talker training, synthetic-talker training, and testing. This sequence of events was similar to that used in Experiment 2.

Familiarization. During the familiarization phase, the subjects were presented with a natural token of each of the four voices in the following sequence. First, a warning light was illuminated to indicate the beginning of a trial. Then the first word from the natural word list was presented over each subject's headphones. After this, a lamp was turned on over button number 1 on the response box to indicate the identity of the talker. On the next trial the same word was presented again, but this time by talker 2, and the lamp over button 2 was illuminated. This sequence was repeated so that one word was repeated successively by each the four talkers before a new word was presented. The sequence continued until all 40 stimuli (10 words by 4 talkers) had been presented. No subject response was required during the entire familiarization procedure. Subjects were told to listen carefully to the words and watch the feedback lights on their response boxes. This phase of Experiment 4 lasted about 5 minutes.

Natural-Talker Training. Natural-voice training was also similar to that used in Experiment 2. The only difference was that the subjects were presented with four talkers rather than six. In the natural-voice training phase, the same 40 stimuli presented during familiarization were now presented random order for identification. After

hearing each word, the subject had to press a button on the response box indicating the identity of the talker. All timing was the same as that in Experiment 2. Four repetitions of each stimulus word were presented for a total of 160 trials. This phase of the experiment lasted about 15 minutes.

Synthetic-Talker Training. Since the subjects would be tested on identification of synthetic stimuli we decided to train them on the synthetic versions of each talker. Each word that was presented was synthesized with the formant pattern, glottal waveform, and the original fundamental frequency contour of one particular talker. The seven words from the word list that were not used in the testing phase (items 3 through 9, inclusive, from Table 4.1) were used for this purpose. The training procedure was identical to the one described above except that four repetitions of each of the 28 stimuli (4 talkers x 7 words) were presented for a total of 112 training trials. This phase of the experiment lasted about 10 minutes.

Testing. After the synthetic talker training phase was completed the subjects were given an extended break of 5 to 10 minutes. Upon their return to the testing room, the upcoming trial sequence was reviewed by the experimenter. Subjects were told that they would hear some of the same words that they had heard in the training phase of the experiment but that the words were changed by computer in various ways. Finally, subjects were told that they would hear one sound per trial and should press the button corresponding to the appropriate talker on their response box. In this phase, however, a new requirement was added to their task. The experimenter explained that after they had identified the talker, they were also to rate the "naturalness" of the word by pressing a button from 1 (very unnatural or mechanical) to 7 (very natural or human). Subjects were also told that at the appropriate times during a trial they would be prompted by the TV monitor at their stations for the appropriate responses that were required.

Since the trial sequence in this phase of Experiment 4 was complex and consisted of several components, a detailed description follows. At the beginning of each trial the word "READY" was displayed on a CRT at approximately eye level for 750 msecs. After a 500 msec pause, one of the 240 unique stimuli was randomly selected and presented to subjects over headphones. At the same time, the instructions "Please identify talker" were centered on the CRT display. An interval of 4 seconds followed during which subjects could respond. At the end of this interval, the CRT display was blanked and a "no response" was recorded for any subject who had failed to respond on that trial. If all the subjects responded before the end of the 4 second interval, the next step in the sequence proceeded early.

After an interval of 500 msec, a new prompt, "Please rate naturalness" was displayed. This began another 4 second interval during which subjects were required to respond. After all the subjects had responded or the response interval had timed out, the CRT display was blanked for a 2 second inter-trial interval before the next "READY" prompt. This sequence was repeated for each of the 240 trials in the experiment.

In contrast to the training phase, the feedback lights and the warning lights on the response boxes were not used here. During this phase only one talker identification response and one naturalness response were collected per stimulus from each subject. In order to collect enough data for a stable measure on each stimulus, a second session was required. This phase of the experiment lasted about 40 minutes.

Procedure -- Day 2

On Day 2, a review of the natural utterances of the four talkers was conducted followed by two more testing conditions. No familiarization or synthetic word training conditions were run. The natural speech review session consisted of a training phase using the 10 natural voice stimuli in a procedure identical to that conducted on Day 1 although the stimulus presentation order was different. This procedure was conducted in order to reacquaint the subjects with the natural voices that they had been exposed to on the previous day.

The testing conditions on Day 2 were also identical to those used on Day 1. The same synthetic words were presented and both identification and naturalness responses were collected. A five to ten minute break was provided between the two blocks of testing. The entire session on Day 2 took about two hours to complete.

Results

The results were separated into the training and testing phases of the experiment. Examination of the data collected during the training phase demonstrated in several ways that subjects had learned the talkers by voice well enough to be tested on synthetic tokens of the same talkers. The results of the testing phase indicated which attributes or components of the talkers' voices controlled the listeners' identification responses. The results from this testing phase also provided information about the relation between the components of talker identity and listeners' naturalness ratings.

Training

The subjects in the present experiment were trained on four talkers rather than six to insure a higher level of talker identification accuracy than we found in the training phase of the glottal identification experiment (Experiment 2). The results summarized below demonstrate that this training procedure was successful for both natural and synthetic stimuli.

Natural Tokens. The left-hand side of Figure 6.1 shows the percent correct identification during the training phase of Day 1 using natural tokens. The right-hand side shows the percent correct obtained in Experiment 2 for the same stimuli (previously shown in Figure 4.1). These two sets of results can be meaningfully compared because the entire procedure including familiarization and training was identical in both experiments. The most noticeable aspect of this comparison is that the performance levels were much higher in the present experiment than in the earlier one [$F(1,37) = 18.14$, $p < .0001$]. The mean percent correct in the present experiment was 92%; the mean percent correct in Experiment 2 was 80%. These comparisons were based on stimuli from talkers that were common to both experiments.

In comparing these training results, it should be kept in mind that reducing the number of talkers from six to four not only made the task easier by reducing the memory load, but it also increased the level of chance responding from 17% to 25%. An information theoretic method of making an unbiased comparison in a similar situation was proposed by Pollack, Pickett, and Sumby (1954). The measure referred to as "percent information transmitted" was defined as the ratio of the information output versus the information input, where each was measured in bits. In their study, Pollack et al. showed that for ensembles of 2 to 16 familiar speakers (1 to 4 bits), the percent of information transmitted was nearly constant. That is, regardless of the set size (within their limits), the voices of talkers that were equally well known received equal scores using this measure.

Since "information" defined in this sense is directly related to the number of choices available, the "information input" was based on the number of talkers in the ensemble, and the "information output" was based on the mean number of talkers chosen correctly. In the present comparison, the information input was always 2.585 bits[1] in the six-talker environment of Experiment 2 and 2.0 bits in the four-talker environment of Experiment 4, and the information output was based on each subject's accuracy level. After this measure

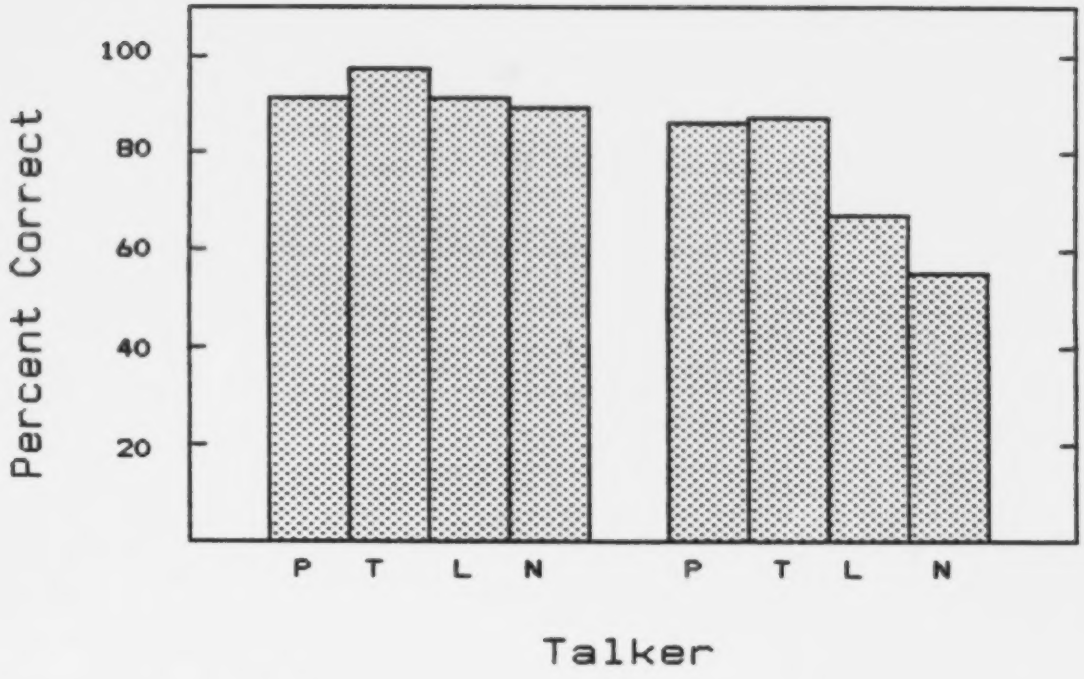


Figure 6.1. A comparison between the natural voice identification accuracies during the training phases of Experiments 4 and 2. The results for Experiment 4 are shown on the left and those for Experiment 2 are shown on the right.

was computed for each listener in both experiments, a t-test showed that the percent information transmitted was greater in this experiment than in Experiment 2 [$t(37) = 4.24$, $p < .0002$]. Therefore, even when a method was used that provided a more sensitive measure of the degree to which a talker's voice had been learned, independent of the simplicity of the task, the subjects were trained to a much higher level of accuracy in the present experiment.

No significant difference in identification performance between the individual talkers was observed in the present experiment [$F(3,36) = 2.79$, $p < .06$]. However, reliable differences between talkers were observed in Experiment 2 [$F(3,75) = 21.67$, $p < .0001$]. The finding that all of the talkers in Experiment 4 were learned to nearly the same degree of accuracy therefore made it possible to interpret the results of the testing phase solely on the basis of the synthesis parameters that were manipulated without having to make a correction based on a priori differences in talker identifiability. Thus, we conclude from these analyses that the training phase in Experiment 4 achieved its goals of insuring good and highly consistent talker identifiability.

Surprisingly, the subjects' identification performance during the training phase of Day 2 (79%) was poorer than it had been on Day 1 (92%) [$F(1,12) = 5.78$, $p < .04$]. This finding can be seen by comparing the right- and left-hand sides of the graph in Figure 6.2. These results were unexpected because the stimuli and the training phase procedures were identical on the two days, and only the random ordering of the trials was different. Two reasons for the differences are suggested by the procedures. First, prior to the training phase on Day 1, a familiarization procedure was used to briefly introduce the subjects to the talkers' voices; on Day 2 this procedure was not used. Second, the testing phase which intervened between the training phases of Day 1 and Day 2 may have had deleterious effects due to the large number of ambiguous or confusing stimuli presented during testing in which the cues for talker identification were permuted from their natural values. In any case, the identification accuracy on Day 2 was considered to be sufficiently high to warrant examining the Day 2 testing data.

Another finding from the Day 2 training condition contributed to a straightforward interpretation of the testing data. The small performance decrement found in the Day 2 training phase did not lead to relative differences in identification performance [$F(3,36) = 2.13$, $p < .11$.]. Combined with the Day 1 training results, this means that all talkers were identified equally well. Another interesting finding follows from this result: One might interpret the absence of any differences in performance between talkers on Day 1 as a ceiling effect; that is, it

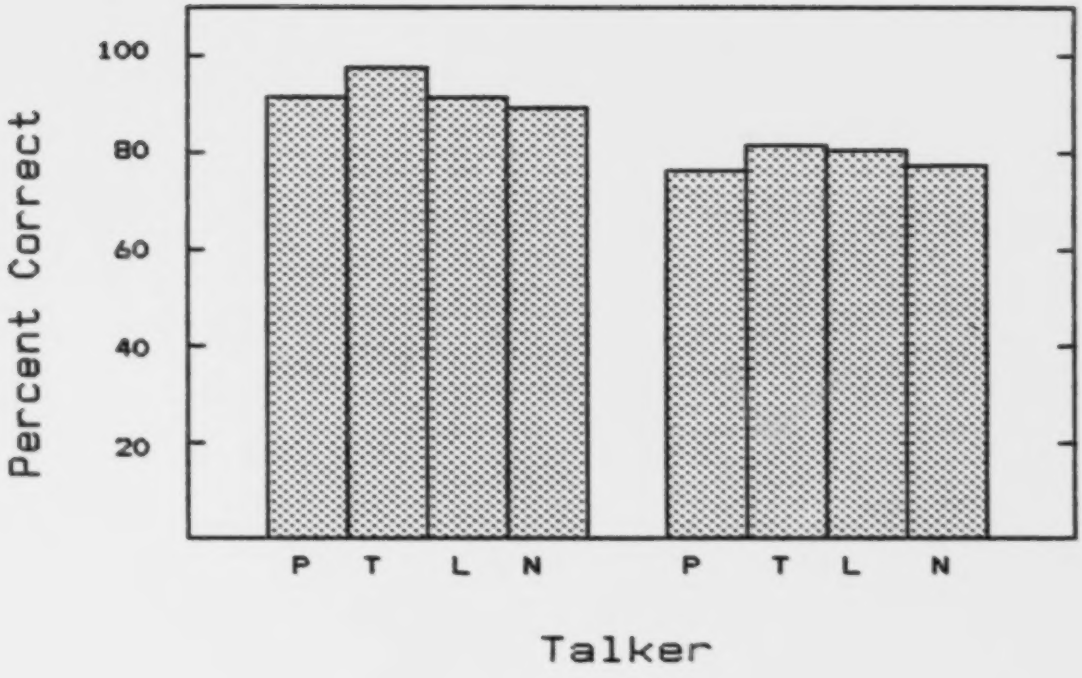


Figure 6.2. Natural voice identification accuracies for each talker measured on Day 1 (left) and Day 2 (right) of Experiment 4.

was possible that real differences in the talkers' identifiability were present but they were not revealed on Day 1 because performance on even the most difficult talkers was so high. The results obtained on Day 2 make this interpretation unlikely and further support the claim that all voices were equally well learned.

Synthetic Tokens. Performance during talker training using the synthetic words was worse than talker training using the natural words although identification was still well above chance. The mean percent correct talker identification was 72%. The performance scores for the individual talkers are shown in Figure 6.3. Small and significant differences in identification between the talkers was observed and confirmed by a 1-way analysis of variance ($F(3,36) = 3.19$, $p < .04$). Since the natural voices were learned equally well, these results indicate that the synthesis techniques did not fully capture the specific attributes equally well for each talker. The magnitude of this effect was quite small with respect to the proportion of variance accounted for (the ratio of the Talker sum of squares to the total sum of squares in the preceding ANOVA was .10) and should not pose serious problems for the interpretation of the experimental manipulations.

Identification Testing

As predicted, neither formant pattern, nor glottal waveform, nor fundamental frequency was the sole invariant attribute used to specify talker identity. Although all of these cues were significantly related to talker identification, none can be singled out as the most salient attribute of the speech waveform. Since 240 interrelated stimuli were used in the present experiment and no truly "correct" responses were appropriate for most of them, the entire pattern of identification results is somewhat overwhelming. Therefore, a simplified analysis of the identification results will be presented before proceeding to a much more detailed analysis.

Our initial data analysis asked the question, "How much does the absence of each individual cue affect overall identification accuracy?" To answer this question, we determined the percent correct when all three cues specified one talker, and compared this value to the percent correct when two of the cues specified one talker and the third cue was disregarded (by averaging across tokens with all possible values of the third cue).

Unfortunately, with the present data it was impossible to legitimately determine the percent correct when all cues specified the same talker. This was due to the fact that

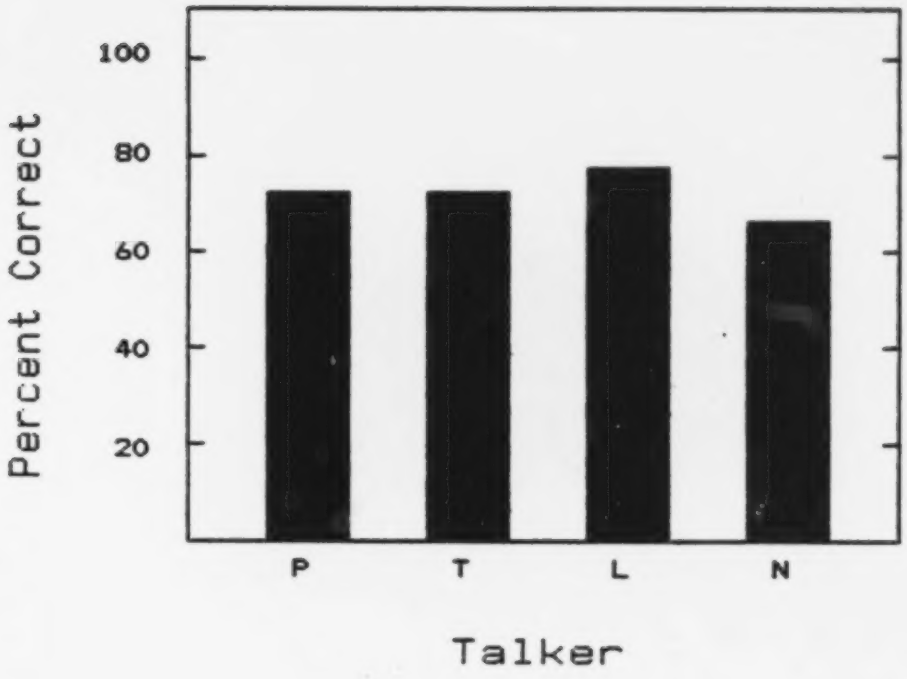


Figure 6.3. Synthetic voice identification accuracy during the training phase of Experiment 4.

five fundamental frequencies were used which covered the range of the four talkers, and no single fundamental frequency corresponded to a single talker. For the sake of simplicity, however, an analysis was conducted which assumed that stimuli with a fundamental frequency of 120 Hz originated from talker P, those with 140 Hz originated from talker T, those with 190 Hz originated from talker L, and those with 220 Hz originated from talker N. These values were the closest matches to the mean fundamental frequencies produced by each of the talkers during the stimulus recording sessions.

Figure 6.4 shows the difference in percent correct between the three cue, and the two cue stimuli. All values below zero reflect a performance decrement due to the missing cue, while all values above zero reflect improved performance due to the absence of the missing cue. The group of bars on the left show the effect of removing fundamental frequency information, the middle group shows the effect of removing glottal waveform information, and the right group shows the effect of removing formant spacing information.

Values above zero might be expected to be small and infrequent since one might assume that any information is better than no information. However, inspection of Figure 6.4 reveals this prediction to be false. The middle group demonstrates that removing the glottal waveform information actually improves identification performance for each talker. This finding suggests that glottal waveform information actually hinders talker identification. While this generalization is correct, it is also an oversimplification. The relationship between glottal waveform and the other cues to talker identity are complex and must be examined in greater detail. The groups on the left and right of Figure 6.4, however, supported prediction that removing cues would interfere with talker identification performance, although there was surprisingly large variation in the left-hand group.

Statistical tests were not conducted on this representation of the data due to the arbitrary definition of percent correct in the present analysis. This initial presentation of the data was intended merely to be a brief summary of the identification results; a more formal analysis is presented below.

The entire set of conditional identification probabilities for each of the stimuli is shown in Appendix 3. Both the examination of this appendix and the initial analysis of the identification results should make it obvious that the data must be summarized into meaningful categories and tested for specific effects in order to understand the experiment as a whole.

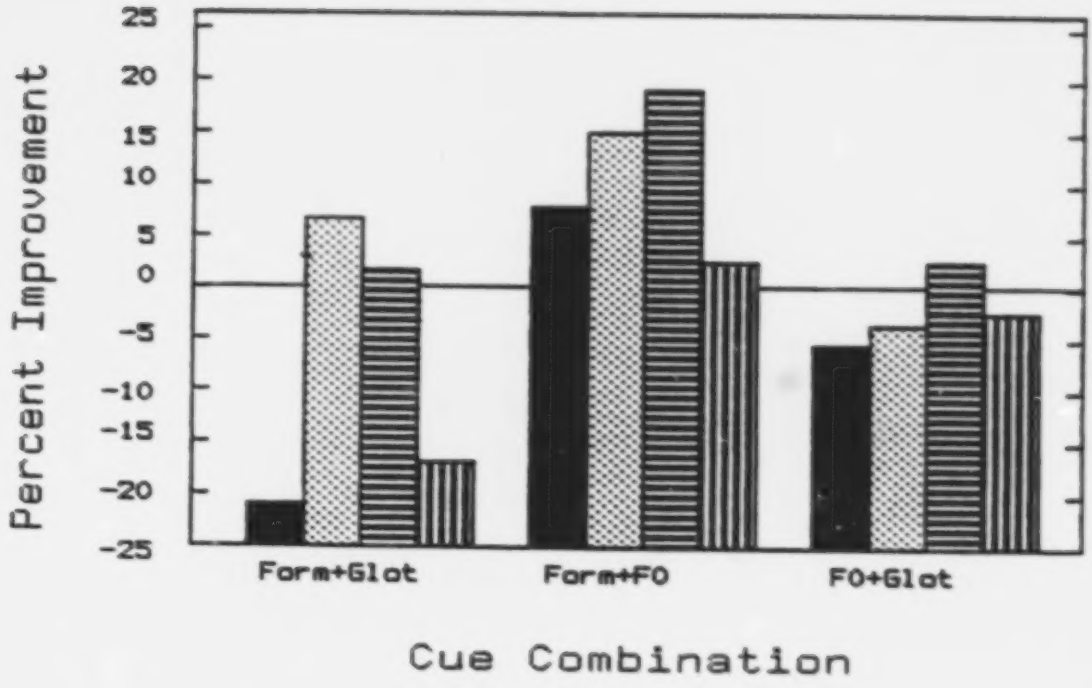


Figure 6.4. Percent change in talker identification due to the lack of fundamental frequency (left group), glottal waveform (center group), and formant spacing (right group). A negative score represents a decrement in performance and a positive score represents an improvement due to the lack of a cue.

To begin this process, fundamental frequency was examined to determine its salience in controlling talker identification. Since fundamental frequency is defined along a simple stimulus continuum, this aspect of the analysis was straightforward. The basic question to be answered was: across all formant spacings and glottal waveforms, are the fundamental frequencies reliably related to both within- and across-gender talker identifiability?

After this analysis, the effects of formant spacing and glottal waveform on talker identification were examined. Since there are no simple physical continua defined for these two potential cues, their analysis was less straightforward. It was necessary to develop two different definitions of a correct response; the first was formant-based and the second was glottal-based. In the first case, a correct response was scored when a subject identified a stimulus as belonging to the talker whose formants it matched, regardless of the fundamental frequency or the glottal waveform of the stimulus. Conversely, in the second case, a correct response was scored (on the same data) when a subject identified a stimulus as belonging to the talker whose glottal waveform it matched. These scores were used to examine both the absolute and relative salience of formant spacing and glottal waveform in talker identification.

Fundamental Frequency. A major finding of the present experiment was that fundamental frequency had a significant effect on subjects' judgements of talker identity. This was observed in spite of the fact that subjects were reminded that in day-to-day conversation, talkers can -- and do -- speak at different pitches. In Figure 6.5 the proportion of responses that were attributed to the male talkers is shown as a function of the fundamental frequency of the stimulus item. Each function represents all the responses to a given talker. The ordinate of each point on the graph is the ratio of the responses for a talker at one frequency to the total number of responses for that talker. This scoring procedure eliminated the effect of response bias towards different talkers and clearly showed that subjects used fundamental frequency in making their choices.

The importance of fundamental frequency to within-gender talker identification was determined in two analyses that examined the male and female response data separately. In the first case, the identification probabilities of the two male talkers showed that lower fundamental frequencies were more likely to elicit P than T identifications. At the lowest frequencies, there were greater proportions of P responses, while at the higher frequencies there were greater proportions of T responses. A strong talker-by-fundamental interaction [$F(4,48) = 24.1$, $p < .0001$]

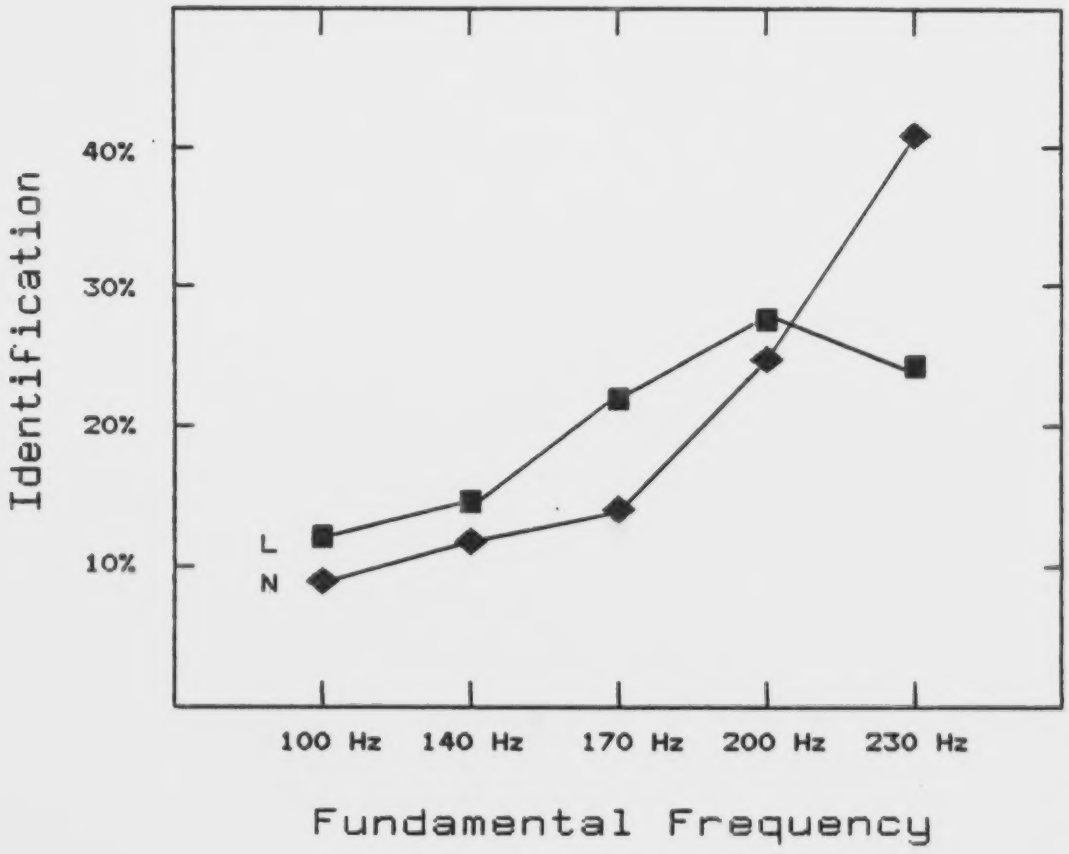


Figure 6.6. Subjects' judgments of female talker identity at five different fundamental frequencies. Responses for talker L are shown with filled squares and talker N with filled diamonds. Each point indicates the percentage of responses for the particular talker at a given fundamental frequency.

supported this interpretation. The analysis of variance also showed a strong main effect of fundamental frequency reflecting fewer P and T responses at higher frequencies [$F(4,48) = 48.8$, $p < .0001$]. This was due to a larger number of female responses at these frequencies.

A similar pattern of results was found in an analysis of the female (L or N) responses. Examination of the response proportions shown in Figure 6.6 shows that stimuli presented at the four lowest fundamental frequencies led to a greater proportion of L responses, whereas stimuli presented at the highest fundamental frequency led to a greater proportion of N responses. These observations were supported by a talker-by-fundamental interaction [$F(4,48) = 15.6$, $p < .0001$]. And, as in the case of the male responses, a main effect of fundamental frequency [$F(4,48) = 89.5$, $p < .0001$] was found. Again, this was due to the fact that the higher the fundamental frequency, the more likely it was that subjects would perceive the voice as a female talker.

Since the relationship between fundamental frequency and talker identification was not entirely invariant it appears that subjects did not rely on fundamental frequency as their only cue in this task. Such an interpretation gains some further support when these data are analyzed in terms of formant spacing and glottal waveform cues.

A comparison between Figures 6.5 and 6.6 also shows a substantial difference in performance across fundamental frequencies between the male and the female talkers. A three-way analysis of variance (sex by talker by fundamental) conducted on this data showed a strong sex-by-fundamental interaction [$F(4,48) = 53.2$, $p < .0001$], indicating that stimuli with lower fundamental frequencies were consistently identified as male (P or T) and those with higher fundamentals were consistently identified as female (L or N).

As noted earlier, the purpose of displaying the data as proportions was to eliminate differences due to talker biases. But it is also important to know whether or not such biases actually exist in the data. In order to examine this, the tests must be based on the raw frequency data, that is, the absolute number of responses to each talker at each fundamental frequency. Data based on frequencies are shown in Figure 6.7 and 6.8 for male and female responses respectively. Such a perspective does not normalize away the absolute differences between the talkers. A two-way (talker identity by fundamental frequency) analysis of variance conducted on the male response frequencies showed no main effect of talker, indicating no overall response bias of one male talker over another [$F(1,212 = 1.07$, $p < .32$]. However, a similar analysis of the female response

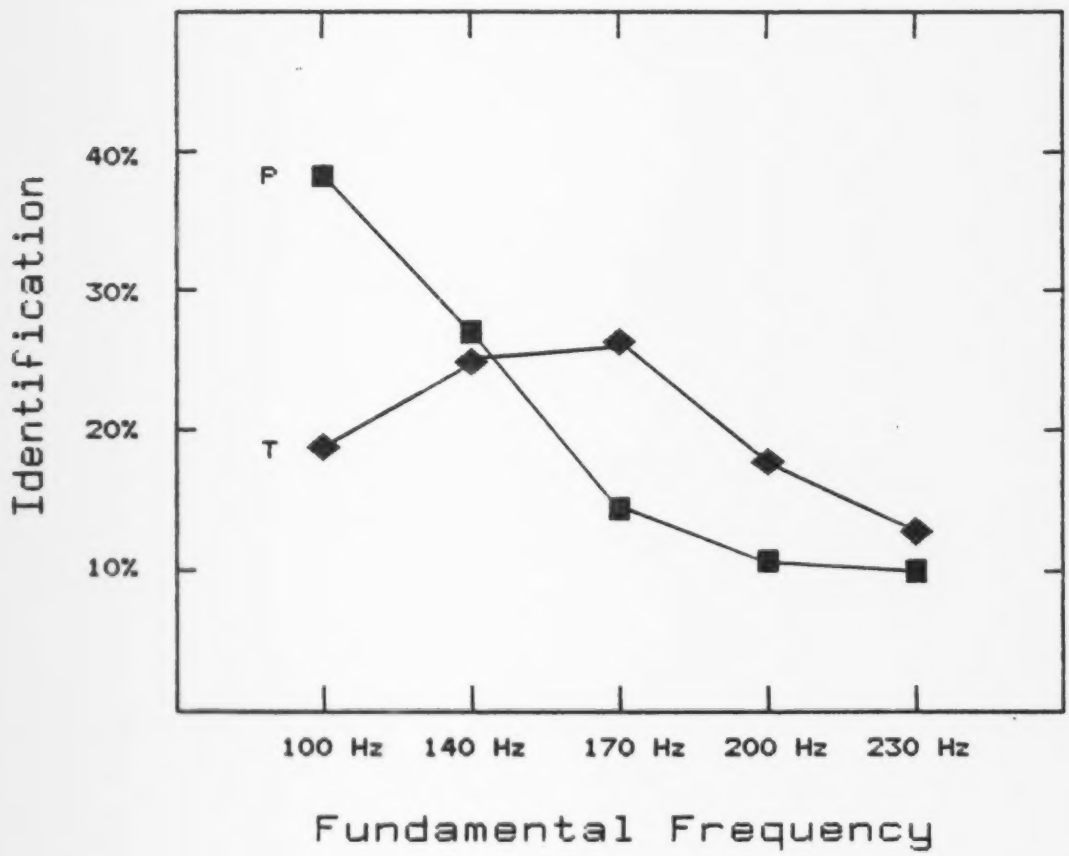


Figure 6.5. Subjects' judgments of male talker identity at five different fundamental frequencies. Responses for talker P are shown with filled squares and talker T with filled diamonds. Each point indicates the percentage of responses for the particular talker at a given fundamental frequency. Responses were averaged across all glottal waveforms and formant patterns.

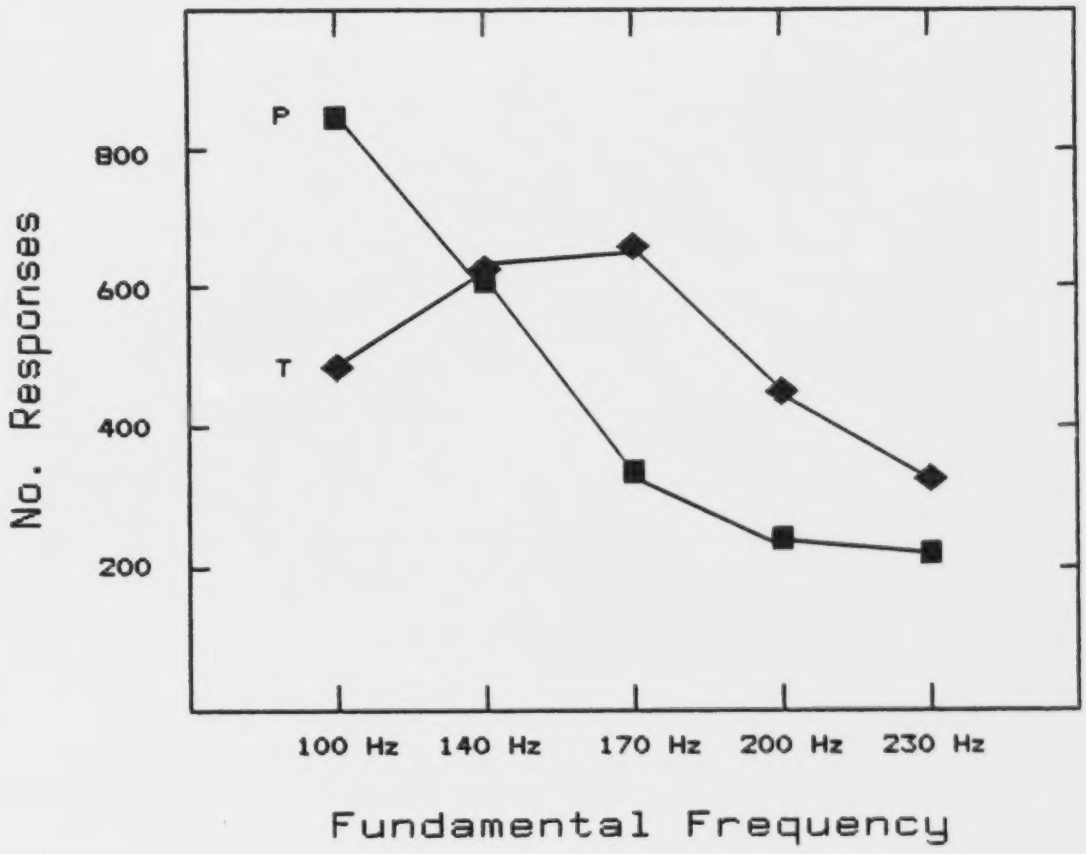


Figure 6.7. Identification frequencies for the two male talkers at each fundamental frequency.

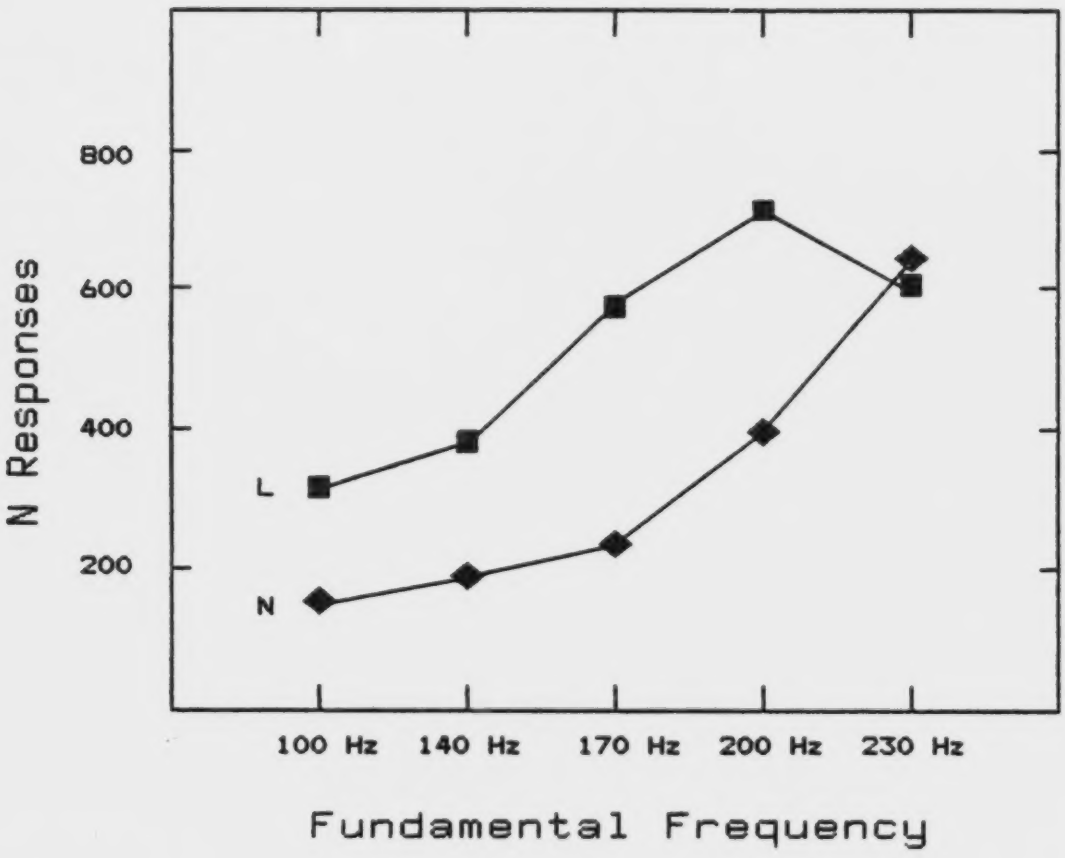


Figure 6.8. Identification frequencies for the two female talkers at each fundamental frequency.

frequencies did show an overall general response bias toward talker L [$F(1,12) = 19.5$, $p < .001$]. One explanation of this finding is that some important cue present in the natural tokens of talker N that the subjects were trained on was not completely preserved in the synthetic versions of her speech. Another is that as stimuli approached the parameters of talker N, the synthesizer failed to produce natural sounding speech due to a breakdown of some of the assumptions of synthesis. The latter possibility will be shown to be unlikely when the naturalness rating data are considered below.

The subjects' responses in this experiment were not only systematic and reliable, but were also based on the fundamental frequency relationships that actually existed between the talkers. The talkers' actual fundamental frequencies, as measured from the voiced portions of the natural words that were presented in the training sessions, are shown in Figure 6.9. By comparing the actual fundamental frequencies with the modal responses for the synthetic stimuli shown in Figures 6.5 and 6.6, it can be seen that subjects used the fundamental frequency information from the words presented in the training phase to identify talkers. Most P responses occurred with stimuli constructed using the lowest fundamental frequency, 110 Hz, whereas most T responses occurred at the second and third lowest levels, 140 and 170 Hz, respectively. Most N responses occurred with stimuli constructed using the highest fundamental frequency, 230 Hz, whereas most L responses occurred at the second highest level, 200 Hz.

Although subjects' responses centered around the fundamental frequencies of the appropriate talkers, many responses occurred not only at adjacent fundamentals but at all fundamentals for each talker. It appears unlikely that this spread of responses could be due to noise or poor training. Indeed, the next two sections demonstrate that formant spacing and glottal characteristics of the talker account for most of the variation observed.

Formant-based Accuracy. Subjects identified the synthetic talkers in a manner consistent with the formant spacing characteristics of each word on 34% of all trials. This value, while significantly greater than the chance level of 25% [$t(202) = 9.1$, $p < .0001$], does not appear to be an especially overwhelming effect. There were, however, environments in which formant spacing played a critical role in controlling talker identity.

One method of displaying the interaction of formant spacing and glottal waveform on talker identification is shown in Figure 6.10. As mentioned earlier, the design of this experiment permitted no correct responses for any of the test stimuli. Therefore, two independent definitions of

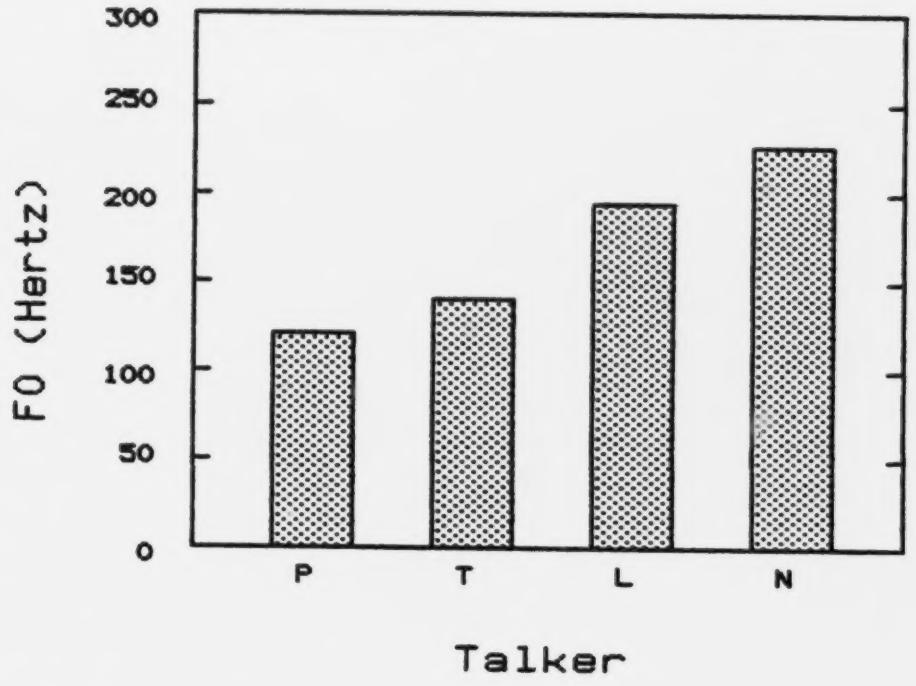


Figure 6.9. Median fundamental frequencies of the natural tokens recorded by each talker.

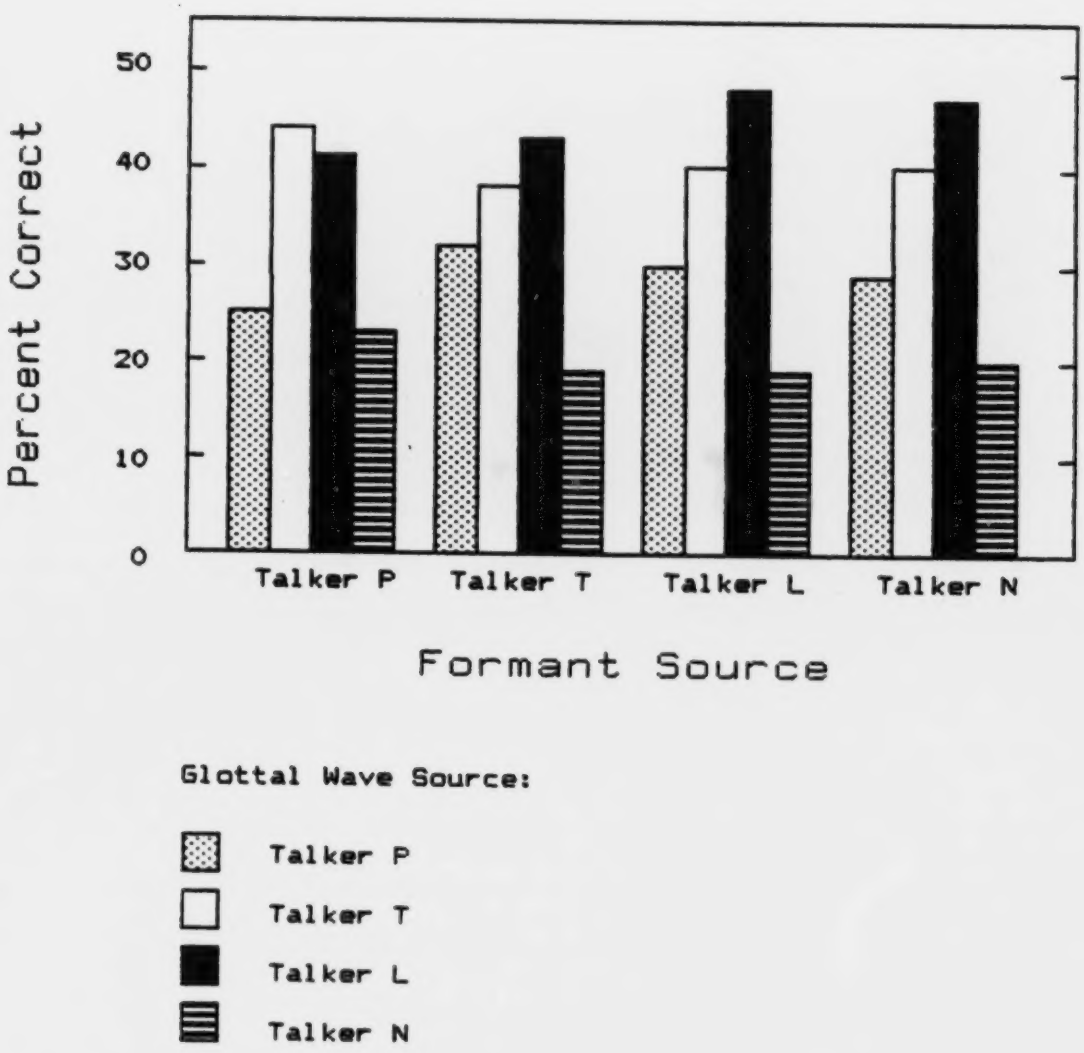


Figure 6.10. Formant-based accuracy for all stimuli showing each combination of formant spacing and glottal waveform. The accuracies are grouped by formant source and the shading of each bar indicates the glottal waveform source.

correct responses were developed. The results of the first, a formant-based measure, are shown here. The vertical axis represents the percentage of stimuli that were identified according to the formant structure of the word, that is, the percentage of stimuli for which subjects responded with the identity of the talker who contributed the formant pattern that was used to generate the word. Each of the bars represents a group of stimuli with a particular combination of formant and glottal cues. The bars are separated into groups of four by the talker who contributed the formant spacing information. The first group consists of stimuli constructed with P formant patterns, the second with T formants, the third with L formants, and the fourth with N formants. The shading of the bar indicates the contributor of the glottal waveform. Thus, the speckled bar in the third group represents the mean formant-based accuracy for all stimuli produced with the formant spacing from talker L and the glottal waveform from talker P. Note that the fundamental frequencies, individual words, and repetitions have been combined in this display so that each bar represents 585 responses.

One might reasonably expect that the highest formant-based accuracy scores would be obtained with stimuli which also shared the glottal waveform taken from the same talker (e.g. the speckeled bar in the first group, the white bar in the second group, the black bar in the third group, and the striped bar in the fourth group). Inspection of the figure reveals that this is not the case. For example, the P formant stimuli were most likely to be judged as belonging to talker P when paired with the T glottal waveform and the N formant stimuli were most likely to be judged as belonging to talker N when paired with the L glottal waveform. A clue to the basis for this finding can be seen in Figure 6.11. This figure shows the same data as shown in Figure 6.10 but now grouped according to glottal waveform rather than according to formant spacing. The first group consists of stimuli constructed with P glottal sources, the second with T glottal sources, the third with L glottal sources, and the fourth with N glottal sources. Note that this configuration clearly shows that some glottal waveforms improved the salience of the formant cues in this task and others reduced this salience. The glottal waveforms of talkers T and L (the middle two groups in Figure 6.11) were significantly better at supporting formant based responses than were the glottal waveforms of talkers P and N. A two-way analysis of variance was conducted on the data shown in Figures 6.10 and 6.11. The results confirmed this finding. While no main effect of formant pattern on formant-based accuracy was observed (in general the formants of all four talkers were equally well identified) [$\underline{F}(3,36) = .56$, $\underline{p} < .64$], a very strong effect of glottal waveform on the same was observed [$\underline{F}(3,36) = 24.93$, $\underline{p} < .0001$]. This supports the argument that some glottal waveforms are better than others for

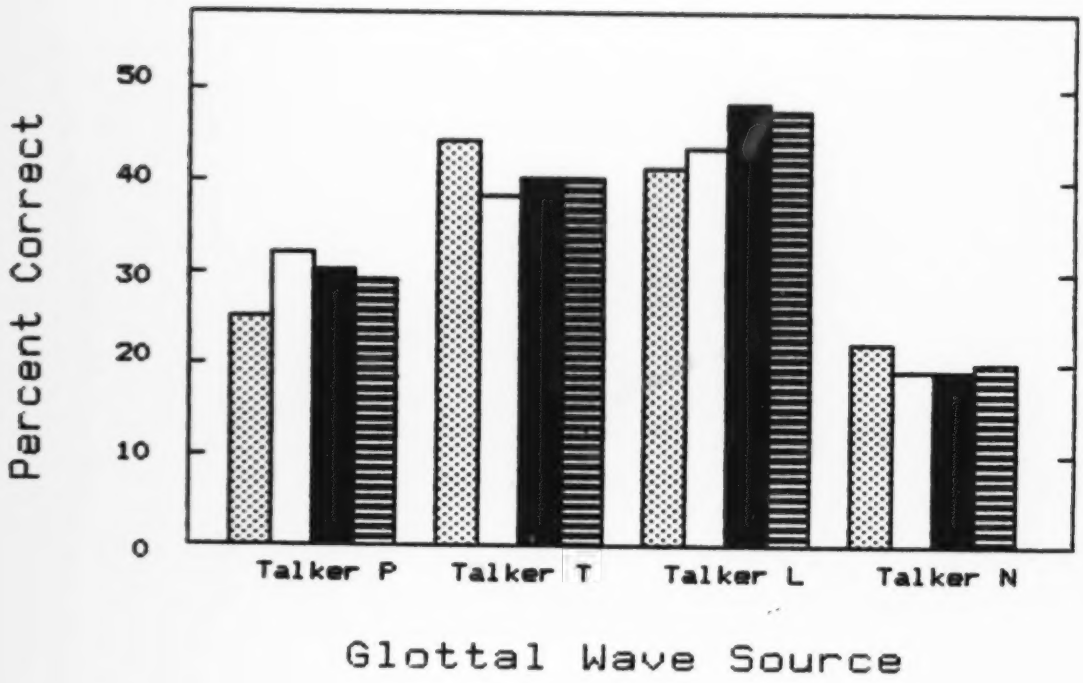


Figure 6.11. Formant-based accuracy for all stimuli showing each combination of formant spacing and glottal waveform. The accuracies are grouped by glottal waveform source and the shading of each bar indicates the formant source.

formant-based identification scores. A formant by glottal interaction was also found [$F(9,108) = 2.91$, $p < .005$]. As shown in Figures 6.10 and 6.11, this reflects the finding that certain glottal waveforms and formant patterns combine to produce higher formant-based identification responses.

Glottal-based Accuracy. Another perspective may be gained by considering the extent to which subjects' responses were controlled by the glottal source cues in the signals. Subjects identified synthetic talkers in a manner consistent with the glottal waveform characteristics of each word on 23% of all trials. This value is slightly below chance, indicating that under these conditions talker identification was not controlled by properties of the glottal waveform. Figure 6.12 shows some details that are illuminating in this regard. The two highest glottal-based accuracy levels were found in those conditions where the stimuli were produced with both the formant and the glottal cues of the same talker (the white bar in the second group and the black bar in the third group). With one exception, the remaining stimuli were identified on the basis of their glottal waveform at levels either at or below chance. Although this data was based on stimuli at all fundamental frequencies, the same pattern of results was found for each one of them taken alone, further supporting the claim that only in cases where both the formant spacing and the glottal waveform were derived from the same talker did responses based on glottal waveform occur at levels above chance.

An analysis of variance was conducted on the glottal-based accuracy scores shown in Figure 6.12. This analysis showed an effect of both glottal waveform [$F(3,36) = 2.99$, $p < .05$] and formant pattern [$F(3,36) = 7.95$, $p < .0003$]. However, the strongest effect in this two-way analysis of variance was the formant by glottal interaction [$F(9,108) = 20.08$, $p < .0001$] reflecting the fact that the best performance occured in conditions where the formant spacing and glottal source characteristics for a talker coincide.

Taken together, the results of the glottal-based identification analysis indicated that listeners used the glottal waveform information to enhance fundamental frequency and formant pattern cues. However, glottal waveform did not appear as an independent source of information about talker identity. This finding is especially important when combined with the results of the formant-based identification measure presented earlier. Recall that in the earlier analysis, formant-based identification was heavily dependent on the particular glottal waveform used in synthesis of the tokens. The dependency was not based on matching the formant spacing of a talker with the glottal waveform of the same talker; rather, two particular glottal waveforms improved performance on all formant patterns. Therefore, glottal

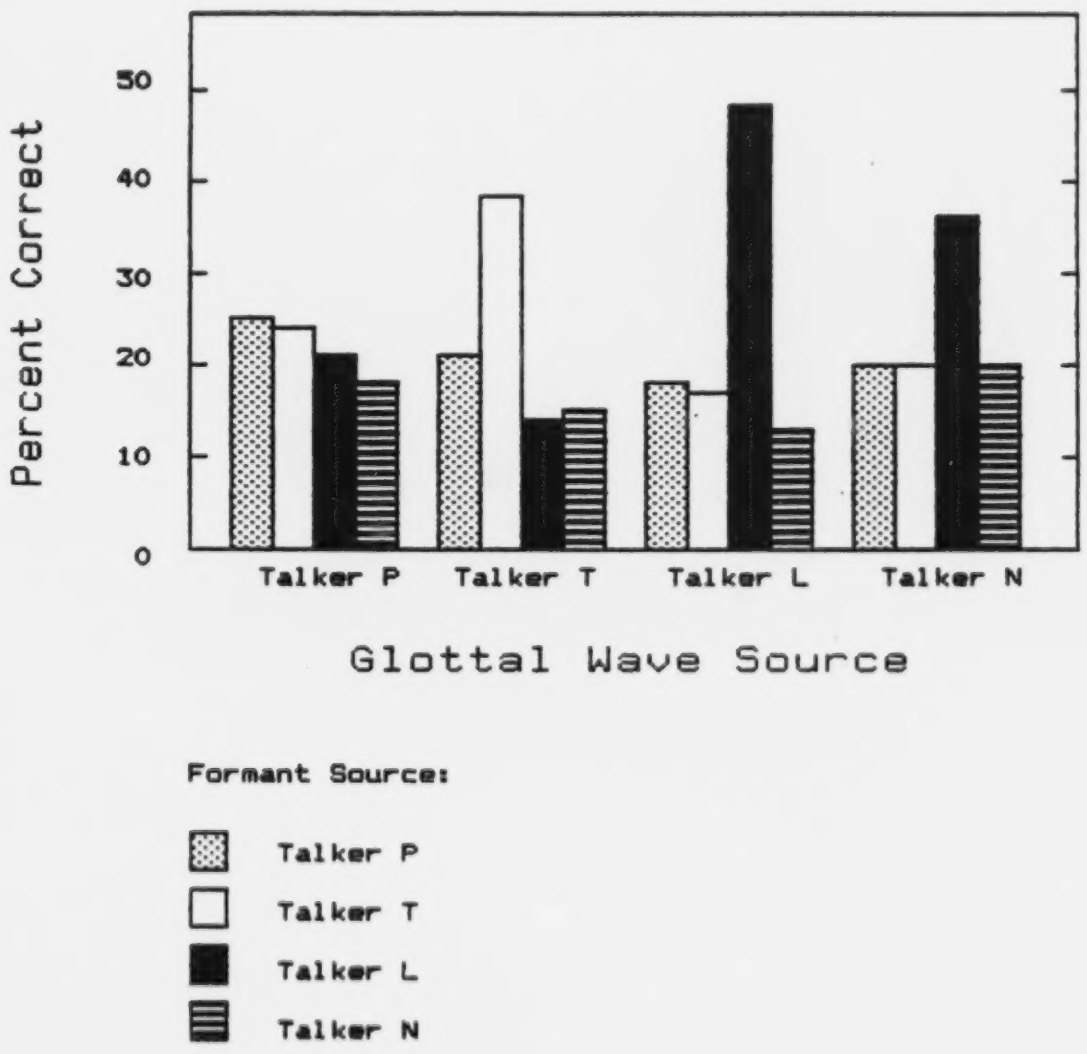


Figure 6.12. Glottal waveform-based accuracy for all stimuli showing each combination of formant spacing and glottal waveform. The accuracies are grouped by glottal waveform source and the shading of each bar indicates the formant source.

source information is important in talker identification, although not in the same manner or to the same degree as fundamental frequency or formant spacing.

Talker Gender Identification

The talker identification data was also analyzed according to talker gender identity. Several different accuracy metrics were developed. Two will be presented here: formant-based and glottal-based. In the case of formant-based accuracy, a response of either P or T (the male talkers) was considered correct if the stimulus was synthesized with the formant patterns of either P or T, and a response of either L or N (the female talkers) was considered correct if the stimulus was synthesized with the formant patterns of either L or N. The overall accuracy level for this analysis was 58.5%, where chance was 50%. As in the talker specific analyses presented earlier, there was much unaccounted for variation due, primarily, to the other two factors; but the overall accuracy on this measure was well above chance [$F(12) = 8.60$, $p < .0001$].

The glottal-based gender identification accuracy measure was constructed in a similar way. A response was judged as correct if the gender of the response corresponded to the gender of the talker who contributed the glottal waveform of the stimulus. And, again, as in the talker specific analyses, no overall effect of glottal waveform on talker gender identification was observed. The mean accuracy using this measure was 48.3%, a result that was not significantly different from chance [$F(12) = 2.08$, $p < .06$].

Naturalness

Naturalness ratings were collected along with talker identifications on each trial of the experiment. We anticipated that the stimuli generated with cues from a single talker would be rated as more natural than stimuli that were generated with the cues combined from several talkers. Surprisingly, Figure 6.13 shows that this was not the case. In this display, the mean naturalness ratings are presented for each talker. Each bar is a particular combination of glottal source and formant pattern. The bars are grouped separately by the source of of the glottal waveform and shaded by the source of the formant spacing. This display of the data shows that rather than matched formant and glottal cues producing the highest naturalness ratings, all stimuli that were synthesized with the glottal source of talker T (the second group from the left) were rated highest. The results of a two-way analysis of variance (formant by glottal) showed a main effect of glottal waveform [$F(3,36) = 19.35$, $p < .0001$] that supported

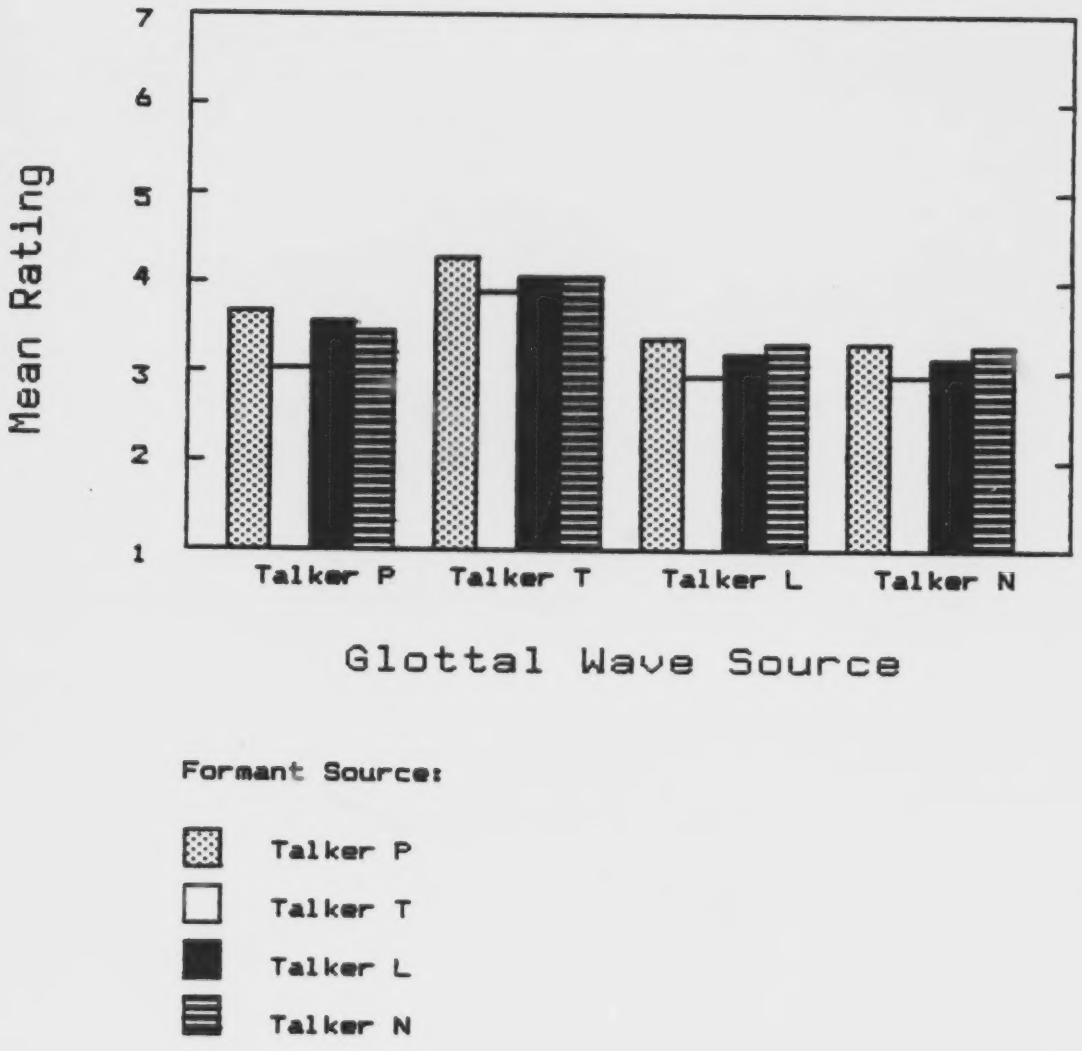


Figure 6.13. Mean naturalness ratings for all stimuli showing each combination of formant spacing and glottal waveform. The ratings are grouped by glottal waveform source and the shading of each bar indicates the formant source.

this observation. Closer examination also revealed that stimuli constructed with talker T's formant patterns were rated lower than those constructed with the formant patterns of other talkers. This result is shown more clearly in Figure 6.14 where the data have been grouped by formant pattern. The mean naturalness ratings in the second group from the left are lower than those in the other groups. This finding produced a main effect of formant [$F(3,36) = 16.06$, $p < .0001$]. A formant-by-glottal interaction was not obtained for these data [$F(3,36) = .96$, n.s.].

Thus, the analyses conducted on the rating data indicate that while the listener's ratings of naturalness are very clearly related to both glottal waveform and formant pattern, these acoustic attributes need not be combined from the same talker.

Reaction Time

Subjects' reaction times for talker identification were also collected and they displayed much the same pattern of results as the naturalness ratings. These latencies were measured from the point at which stimulus output was initiated to the point at which the subject pressed one of the four buttons to identify the talker. Since there were no correct responses in this task, all reaction times were included in the present analysis. Figure 6.15 shows the mean reaction times in the familiar configuration grouped by glottal waveform. Note that the fastest reaction times were obtained from the stimuli constructed with the T glottal source (the second group from the left) and the longest reaction times were obtained from the stimuli constructed with the N glottal sources. A main effect of glottal source found in a two-way analysis of variance showed that the glottal sources were significantly related to reaction time [$F(3,36) = 11.16$, $p < .0001$]. This pattern of results was identical to that obtained with the subjects' naturalness ratings. Specifically, glottal sources that resulted in higher naturalness ratings, also resulted in faster reaction times.

No significant relationship was observed between formant spacing and reaction time [$F(3,36) = 1.17$, n.s.]. And, no interaction was observed between the formant and glottal factors [$F(3,36) = 1.35$, n.s.].

On the suspicion that reaction time and naturalness ratings were providing the same information, a correlation between the two measures was conducted. A significant correlation would be important for a number of reasons, the least of which is that in future experiments it might only be necessary to collect reaction time measures on the identification responses and thereby get naturalness

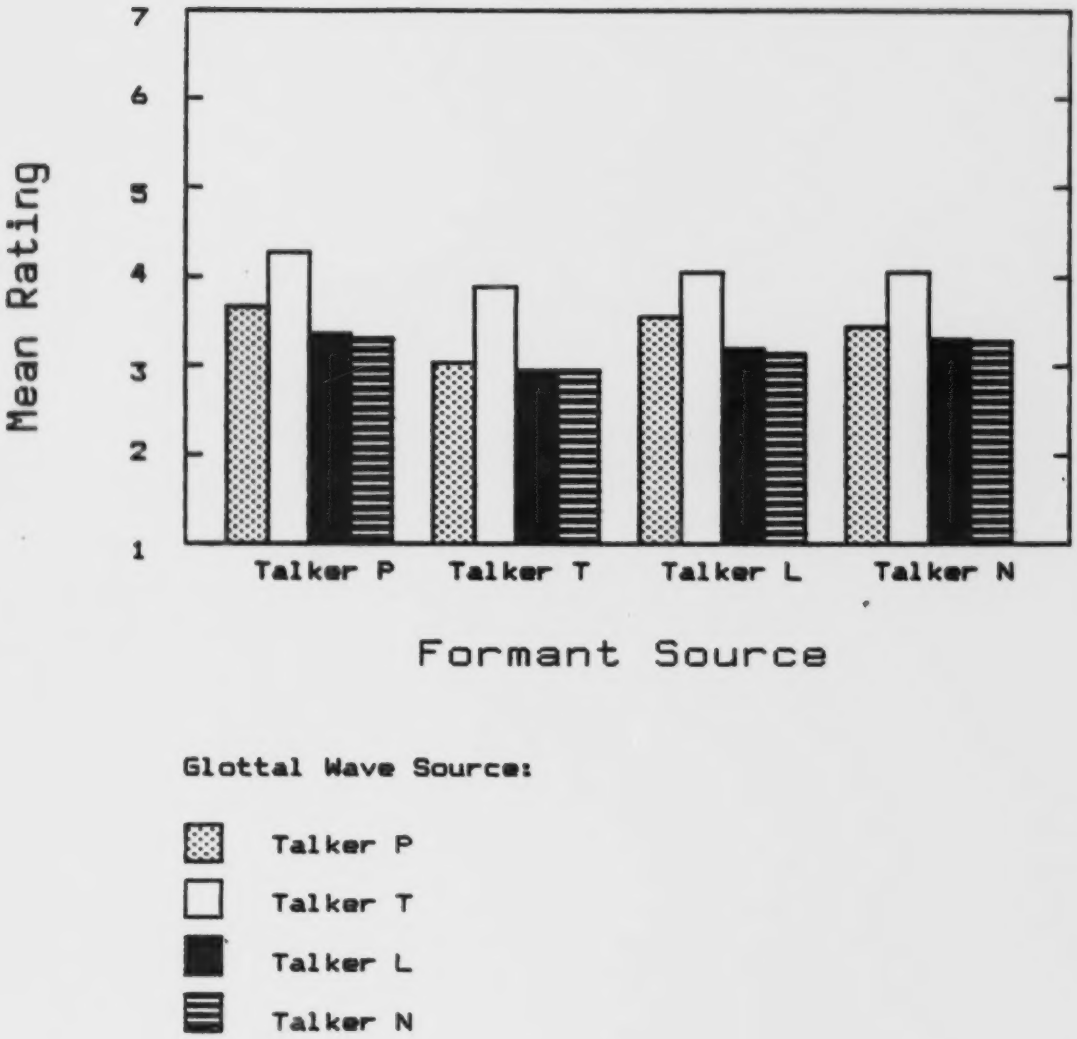


Figure 6.14. Mean naturalness ratings for all stimuli showing each combination of formant spacing and glottal waveform. The ratings are grouped by formant source and the shading of each bar indicates the glottal waveform source.

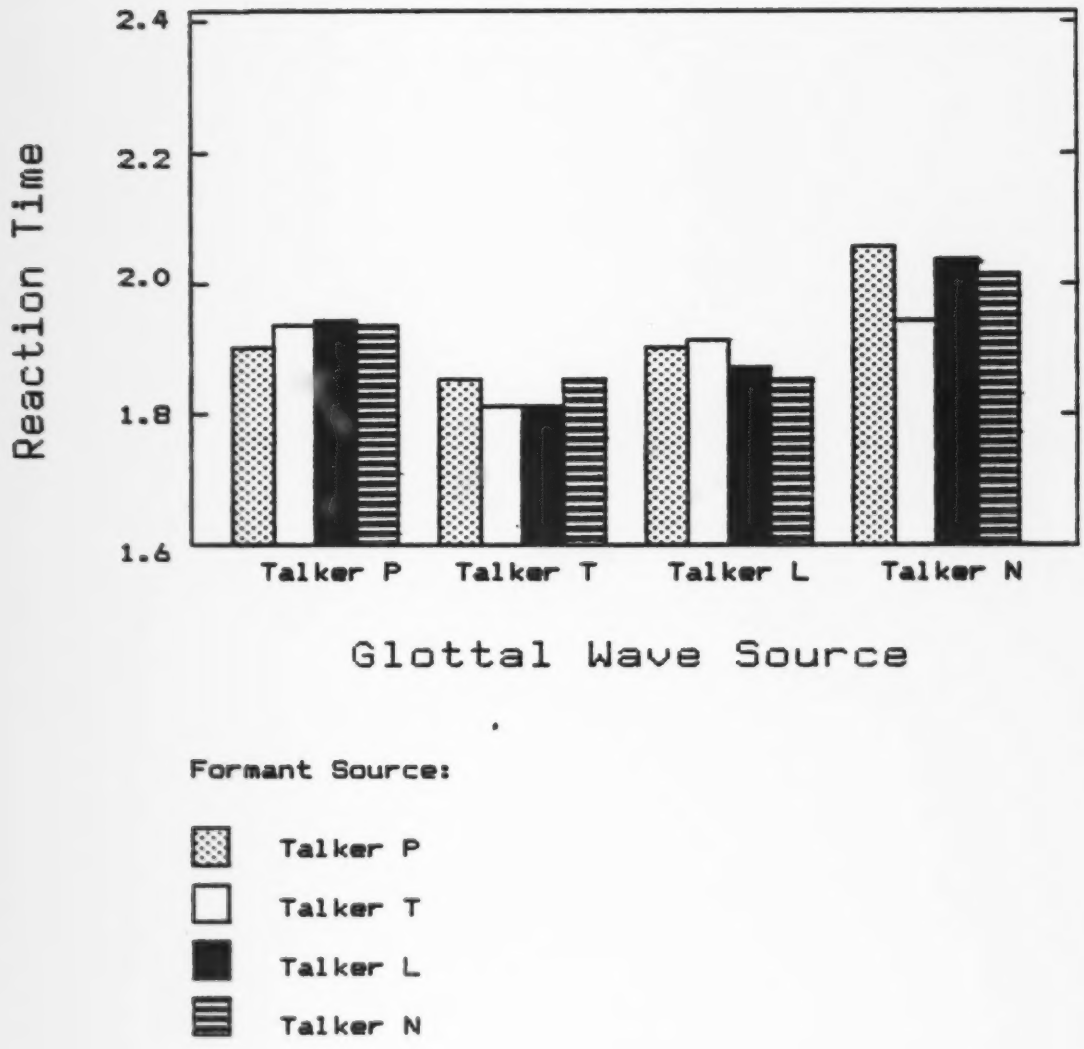


Figure 6.15. Mean reaction times in seconds for all stimuli showing each combination of formant spacing and glottal waveform. The reaction times are grouped by glottal waveform source and the shading of each bar indicates the formant source.

information without collecting separate naturalness ratings. The analysis produced a low and nonsignificant correlation between these two indices [$r = -.079$, $t(206) = 1.14$, $p < .26$]. This result suggests that reaction time in identification should not be substituted for naturalness ratings despite the similarity in the pattern of the data shown here.

Discussion

The present findings

As predicted, fundamental frequency, formant spacing, and glottal waveform were each shown to be important factors controlling the identification of talkers. The manner in which these attributes interacted, however, turned out to be very surprising. We expected that stimuli synthesized with the particular formant spacings, fundamental frequencies, and glottal waveforms of the same talker would be identified more often as that talker; this, however, was not the case. Rather, we found that the ability of formant spacing to determine talker identity depended on the particular glottal waveform it was combined with. The glottal waveforms taken from some talkers supported high levels of formant-based identification for all formant patterns, but those taken from other talkers lowered the formant-based responding, even for those formant patterns that were produced by talkers who actually contributed the glottal waveform.

The best independent predictor of listeners' judgments of talker identity was formant spacing. However, the performance level using only this cue was still rather low overall (34 percent correct) without the additional, redundantly specified information of fundamental frequency and glottal waveform. The fundamental frequencies that were used in this experiment did not exactly match individual talkers (as did the formant spacings and the glottal waveforms); rather, they spanned a range of values produced by talkers in equal steps. But more identifications for a given talker were reported nearer that talker's average fundamental frequency than further from it. Glottal waveform was important in talker identification, but only insofar as it contributed to formant-based identifications; taken alone as an invariant cue to talker identity, glottal waveform contributed only at chance levels of performance.

Since the glottal source played an important role in talker identifiability across different talkers and even across different genders, and since some glottal sources were better than others regardless of the other talker

identity cues, it is important to know precisely what acoustic attributes of the glottal source lead to improved talker identification. Unfortunately, with only four glottal sources, the present investigation did not have a large enough sample size to answer this question definitively. An examination of the Fourier spectra of the glottal waveforms that were used in the present experiment did reveal that neither the slope of the glottal spectrum nor the specific frequencies of the harmonics seemed to underlie the effects that were found. At this point, we can only speculate that the glottal waveforms used in the present experiment were differentially effective in talker identification, and that the common attributes shared by the "good" glottal sources remain to be extracted in future analysis and synthesis studies using a larger number of glottal waveforms.

Several other findings were obtained in the present investigation. The training phase showed that listeners could be trained to identify previously unknown talkers by voice alone, using only a list of monosyllabic words, in a very short period of time (about 20 minutes). Nearly all previous studies have used either familiar talkers in identification tasks or unknown talkers in discrimination tasks. While useful, earlier methods have a number of limitations. Studies using familiar talkers do not have the ability to control talker familiarity, and, in a practical sense, it is very difficult to find a large sample of subjects that know a particular set of talkers equally well. Discrimination tasks, on the other hand, use arbitrarily large subject pools, but interpretation of results is often difficult because of unknown and uncontrolled subject strategies in discrimination. The use of voice identification training as demonstrated in the present experiment will permit the design of a wide variety of new experiments.

The performance of subjects in identifying synthetic talkers by voice was also an important finding because of the synthesis methods used. Recent studies have demonstrated that the Klatt software synthesizer can adequately model speech production in terms of segmental intelligibility (as the final stage of the MITalk text-to-speech system, Pisoni & Hunnicutt, 1980) but its ability to model the talker specific characteristics of speech necessary for the perception of talker identity has never been examined before the present investigation. In the present study, listeners who were trained to identify talkers by voice at a 92% accuracy level were able to identify the synthetic models of these words at an accuracy level of 72% when the same talker identification task was used for both measures. While the synthetic speech revealed significantly lower accuracy scores, the levels were far above chance and show that much of the information necessary

for talker identification was, indeed, preserved by the synthesis methods in terms of the three parameters that were manipulated. Therefore, 72% is a lower limit on the ability of the modified Klatt synthesizer to preserve the acoustic correlates of talker identity. No doubt better modelling of more talker-specific cues would improve this figure.

The overall identification and naturalness results showed that the modified Klatt synthesizer preserved perceptually important talker differences at least moderately well, however, the difference in performance between the natural versus synthetic tokens was substantial and requires an explanation. One possibility is that basic limitations of the synthesizer created a somewhat poor model of the speech production process. Another possibility is that a failure to precisely measure and specify the parameters of interest, in preparation for synthesis, might have created this difference. Finally, it might be the case that the three parameters that were manipulated only partially specified perceptually important talker differences.

The problem of determining whether it was missing cues or poor synthesis that was responsible for the accuracy differences between the natural and synthetic versions of the talkers' words may be approached from two directions. The possibility of a poor synthesis model (with respect to talker identification) could be discounted if listening tests showed that the systematic addition of cues differentiating between talkers to the synthesis parameters improved synthetic talker identity to natural levels. On the other hand, a more fine grained synthesis system, such as a high-bit-rate LPC vocoder, could be used to approach this problem from the other direction. Very precise models of the speech production process produced by such a device could be degraded to remove certain acoustic cues from the waveform. In terms of the parameters studied in the present investigation, it would also be necessary for this system to be equipped to accept arbitrary, glottal source information.

The naturalness ratings that were collected were similar in several ways to the talker identification results. As in the identification task, main effects were observed for formant spacing and glottal waveform, but the highest naturalness ratings were not produced by the synthetic speech that duplicated those aspects of particular natural talkers. Some talker's glottal sources and, to a somewhat lesser extent, some talker's formant spacings produced more natural sounding synthetic speech regardless of similarity of the combination of cues to particular natural talkers. Once again, there were too few glottal waveforms to determine precisely what aspect related to naturalness. We can only conclude at this point that the particular natural glottal waveform used in stimulus

synthesis has a very strong effect on the naturalness ratings of synthetic speech.

A comparison with the first three experiments

In spite of its limitations, and in addition to the results already cited, the present experiment also provided a new perspective from which the first three experiments can be analyzed. For example, in Experiment 1, it was found that those stimuli synthesized with a male glottal waveform were perceived as male more often than those synthesized with a female glottal waveform. The results of the present experiment showed no effect of glottal waveform on the perception of talker gender. Although the task in this experiment was not specifically talker gender identification, the error data was analyzed in terms of talker sex and it was found that glottal waveform was not related directly to talker sex identity. Given this result, the earlier finding is somewhat surprising. Although the tech niques used in Experiment 1 were probably more sensitive than those used in Experiment 4, the large magnitude of the glottal waveform effects found among the four talkers in the present experiment suggest that the earlier results were specific to the two glottal waveforms that were chosen. By combining the results of both experiments, we conclude that while some glottal waveforms are reliably and consistently related to talker gender identification (as evidenced by the two waveforms used in Experiment 1), this should not be taken as a general rule. Additionally, these contrasting results reemphasize the importance of selecting a larger set of talkers in experiments that extend the present work.

The results from Experiment 2, the glottal identification experiment, are also interesting in light of the results from the present experiment. In particular, the talkers who were used in the latter experiment were a subset of those used in the former experiment, making the results more comparable. The combined results from both experiments indicated that even though listeners were able to use the glottal waveform and associated waveform in isolation to identify talkers, the glottal waveform was at best only indirectly useful in talker identification after being filtered by the vocal-tract transfer function.

The naturalness results from the present experiment allowed a more definitive interpretation of some of the results found in Experiment 3. Recall that the naturalness ratings were lower for female talkers than for male talkers in Experiment 3. Earlier we suggested that this was due to the fact that the natural voices of the female talkers were not learned as well as the natural voices of the male talkers. This explanation was not supported in

Experiment 4. The male and female talkers were learned equally well in the present experiment yet the naturalness ratings for those stimuli with parameters in common with the stimuli of the earlier experiment were higher for the males than for the females. Furthermore, this also held true across fundamental frequencies when the formant spacings and the glottal waveforms were held constant for the same talkers. Such findings indicate that the lack of naturalness for female speech was not simply due to the synthesizer performing more poorly at higher fundamental frequencies. The particular formants and glottal waveforms of the female talkers apparently resulted in lower naturalness ratings.

Summary and Conclusions

The results obtained in this experiment demonstrate that of the three attributes that were manipulated in a factorial design, formant spacing and fundamental frequency were primarily responsible for contributing to talker identity. Glottal waveform appears to contribute to talker identification only indirectly in terms of the perceived salience of formant spacing in identification. However, glottal waveform is also directly related to measures of perceived naturalness.

The training and testing techniques used here showed that identification of talkers could be studied experimentally using the same analysis and testing methodologies that have been successfully used in the examination of questions related to segmental phonetic perception. It is hoped that with continued experimentation, other attributes of talker identification will allow this indexical property of the speech signal to be described by a systematic set of rules that will capture dialectical, gender, and emotional qualities of a talker.

Footnote

1. The information in bits is calculated as the logarithm to the base 2 of the number of choices available.

Chapter 7

Summary and Conclusions

The experiments described in this investigation examined the effects of fundamental frequency, formant spacing, and glottal waveform on the identification of talkers. The purpose of this study was to gain a better understanding of the acoustic attributes used to perceive talker identity and to provide data for improvment in the quality and flexibility of speech synthesis systems. The specific relationships that were found have been discussed in detail in earlier chapters and will be reviewed here only briefly. Then some of the general conclusions will be outlined.

Experiment 1 was carried out to determine whether listeners' perceptions of talker gender identity were affected by the source spectrum generated by the glottal waveform. A sensitive perceptual technique using an identification procedure was used to assess the importance of glottal waveform. Using this technique, stimuli were constructed which ranged from male to female in terms of formant spacing. Half of the stimuli were generated with a glottal waveform taken from a male talker and half were generated with a glottal waveform taken from a female talker. The results showed that the crossover point of the talker gender identification function depended on whether the glottal waveform had been produced by a male or a female talker. Specifically, in the gender identification tests, listeners reliably switched from "male" responses to "female" responses further along the formant spacing continuum when the stimuli were constructed with male glottal waveforms rather than with female glottal waveforms. The inverse relationship was also observed with stimuli synthesized using female glottal waveforms.

Experiment 2 was designed to extend these findings in order to determine whether the identity of an individual talker could be determined by human listeners from information in the glottal waveform alone. In this experiment, listeners were first trained to identify three male and three female talkers; they were then tested to determine if they could identify these talkers based on glottal waveforms that had been extracted from the speech of each talker. The results showed that listeners were, in fact, able to identify talkers that they had learned well during the training phase of the experiment. They responded at chance levels to the talkers that they had not learned well.

Experiment 3 was performed to evaluate the suitability of the synthesis methods, the Klatt software synthesizer, and the parameter estimation methods, all of which were to be used in Experiment 4. Subjects rated the naturalness and intelligibility of stimuli that were designed to mimic the specific acoustic correlates of three male and three female talkers. The mean ratings were good. The scores generally fell within the upper half of the possible range of values. Naturalness ratings and intelligibility ratings were not correlated, indicating that these two perceptual measures assesed two different qualities of the speech waveform.

Having determined that glottal waveform information could be used to identify talkers and that the synthetic stimuli were acceptable to listeners, Experiment 4 examined the combined effects of fundamental frequency, formant spacing, and glottal waveform on talker identification and ratings of perceived naturalness. A set of stimuli were synthesized based on a factorial combination of these three cues taken from each talker. Listeners were first trained to identify two male and two female takers by voice. When this was completed, listeners were required to identify a specific talker for each of the factorial combination stimuli. Most of the stimuli had characteristics of the voices of more than one talker. The results of Experiment 4 demonstrated that formant spacing and fundamental frequency were the primary sources of information used by listeners in perceiving the identity of a talker. Glottal waveform played only an indirect role in the specification of talker identity. Certain glottal waveforms produced better formant-based accuracy, and other glottal waveforms produced worse formant-based accuracy. However, glottal waveform played no direct role in controlling listener perception of talker identity independent of the other two attributes. Naturalness ratings were also collected for each of these factorial stimuli. The ratings showed that glottal waveform was directly and systematically related to measures of naturalness. Certain glottal waveforms were consistently rated more natural across all fundamental frequencies and formant spacings than other glottal waveforms.

On the surface, at least, Experiment 4 appeared to contradict the results of Experiments 1 and 2. In the first two experiments, glottal waveform was shown to be an important cue for talker identification, whereas in Experiment 4, glottal waveform was shown to be only indirectly useful. This apparent discrepancy highlights the fact that while talker specific information is present in the glottal waveform, other cues obscure its importance in word length utterances. Experiments 1 and 2 used very sensitive labeling tasks that assesed what listeners can do in low uncertainty testing situations, not what they actually do when listening to meaningful words.

The findings of this investigation clarified some of the confusion found in previous studies on talker identification. The results of earlier studies were inconsistent about the precise relation between the characteristics of the glottal waveform and the perception of talker gender. Lass, Hughes, Bowyer, Waters, and Bourne (1976) used whispered and low-pass filtered speech to examine superlaryngeal and glottal cues respectively and concluded that glottal source information was more important to gender identification than vocal-tract resonance information. In a later study, Monsen and Engebretson (1977) showed systematic differences between the glottal sources of male and female talkers independent of fundamental frequency. Their results established that the glottal source might be valuable perceptually and could account for the results reported by Lass et al. Other studies, however, showed that glottal waveform was relatively unimportant in gender identification. Schwartz and Rine (1968), using whispered speech, and Coleman (1971 & 1976), using an artifical larynx, obtained results suggesting that the formant structure was the most important cue to gender identification.

Coleman's conclusions proved to be the most accurate in light of the results obtained in Experiment 4. Listeners were unable to correctly identify the talker's gender from glottal source information alone. While these conclusions appear justified in light of the results, it should be noted that only four glottal waveforms were used in this experiment. Different findings might be obtained by using the glottal waveforms of a much larger number and wider variety of talkers.

In addition to gender identification, several previous studies provided mixed reports about the effect of glottal source on individual talker identity. The work of LaRiviere (1975) indicated that the glottal source and the formant spacings were more or less equivalent for purposes of talker identification. This conclusion was based on a talker identification task in which the stimuli were four isolated vowels produced by eight male talkers. Identification performance was the same for both whispered and low-pass filtered vowels, which conveyed superlaryngeal and glottal information, respectively. On the other hand, Coleman (1973) showed that glottal waveform information was quite unnecessary, at least in connection with stimuli that were five seconds in duration. With the aid of an artificial larynx, his findings showed that excellent (90%) talker discrimination was possible with no glottal source information whatsoever.

Again, these studies used entirely different methodologies from one another and the results were difficult to

compare without additional information. The results of the present investigation provide support for the assertion that the part of the glottal source information that was useful in the LaRiviere experiment was the fundamental frequency and, furthermore, that talker identification is more strongly related to formant spacing than to glottal waveform at least for word length stimuli. The high level of performance in the Coleman experiment was probably due to additional prosodic and timing cues found in his longer-duration stimulus items and the fact that he employed a discrimination task rather than an identification task.

As in previous work, the present investigation found a close connection between talker identification, formant spacing, and fundamental frequency. It should be pointed out, however, that the talker identification accuracy level (based on synthetic stimuli that left both of these cues intact) was still well below performance with natural tokens of the same words. In the training phase of Experiment 4, the identification accuracy was 72% for stimuli that retained formant spacing, fundamental frequency, and glottal waveform information from the original talker, whereas the accuracy level for talker identification for the natural tokens was 92%. While it is clear that formant spacing and fundamental frequency were useful in talker identification, it is equally clear that other important characteristics of the speech were left outof the synthetic stimuli. Moreover, since the natural tokens were only of word duration it is probably safe to say that longer duration prosodic information did not account for the discrepency between the identification accuracy of natural and synthetic speech. More acoustic-phonetic information about talker identification appears to be present at the word level than was accounted for only by the three parameters that were manipulated. It is possible that fine temporal and allophonic differences in phonetic implementation rules also contribute to talker identification.

Naturalness ratings from the stimuli that were used in the identification tasks were examined in Experiments 3 and 4. While Experiment 3 showed that the mean ratings were in the top half of those possible, there was much room for improvement. Furthermore, those stimuli in which all three cues specified one talker were rated no more natural than the mean of all stimuli constructed with all possible combinations of cues. This result indicates that a particular stimulus does not have to sound like a particular person to sound natural. In fact, one parameter was closely associated with talker naturalness. The stimuli that were synthesized with two particular glottal waveforms were rated more natural than the stimuli synthesized with the other two glottal waveforms. This finding held true across all formant spacings and fundamental frequencies.

What aspect of the glottal waveform created more natural sounding speech? Rosenberg (1971) addressed this question directly in his study of the effects of various glottal pulse shapes on listeners' preferences for synthetically produced syllables and sentences. One natural glottal waveform and six synthetic ones were presented to listeners for evaluation. As might be predicted, the natural glottal waveform was rated higher than the artificial ones in a syllable rating task. The synthetic glottal waveforms were constructed from a number of simple mathematical functions that specified the opening and closing phases of the glottal pulse and led to a systematic pattern of preferences. Glottal waveforms containing a single slope discontinuity at closing were rated higher than the others. Although the best natural glottal waveform in the present experiment, as determined by naturalness rating, did fit this description, so did the poorest; tests with additional glottal sources, specified in both the time and frequency domain, will be necessary to extend Rosenberg's results to natural glottal waveforms. Even without a detailed understanding of the characteristics of the glottal waveform that lead to high naturalness ratings, it appears that ratings of naturalness are related to particular stimulus parameters rather than particular talkers.

In summary, the results of the present investigation demonstrate that the shape of the glottal waveform is indirectly related to the perception of talker identity and directly related to perceived judgements of naturalness. The specific characteristics of the glottal waveform that underlie these effects have not been identified at the present time. Fundamental frequency and formant spacing have been shown to be directly related to talker identification, but not to judgements of naturalness. Further research will be necessary to find the specific acoustic attributes of the glottal waveform that led to the interactions in perception that were observed between glottal waveform, fundamental frequency, and formant spacing as cues to talker identification.

Future Directions

It is a simple matter to describe the work necessary to remove the major limitations of the present study. The additions would be: more talkers, more words, more training, and more potential cues. Fortunately, the findings that have been presented here were both interesting and reliable enough to indicate that such extensions would be worthwhile.

With respect to the number of talkers, only two males and two females were used as models in Experiment 4. And,

while four different voices were sufficient to make general statements about the importance of different cues for talker identification, more detailed questions regarding the relationship of the acoustic attributes of glottal waveforms and formant-based identification can only be answered with a larger sample of talkers. This larger sample would entail a longer training period than was provided in the present experiment in order for the listeners to reach asymptotic levels on each of the talkers.

Regarding the number of stimuli, only three different words were actually used in the testing phase of Experiment 4 (although there were 240 unique stimuli). These three words were selected because they contained a variety of phonemes, but obviously, three words cannot represent all possible speech sounds that might be used to differentiate talkers. In order to extend the generality of the findings, additional words should be tested. Such changes would probably also entail a restructuring of the procedure from a within-subjects to a partially between-subjects design.

Finally, as indicated above, more potential acoustic correlates of talker identity should be manipulated in order to explain the difference between natural and synthetic talker identification levels found in here. Previous research has indicated that these other potentially important cues might include timing, both within and between phonetic segments, diphthongization of vowels, generation of fricative spectra, and specific nasal characteristics. Although only mentioned briefly, it would be worthwhile to study the ways in which talkers use low-level phonetic implementation rules to realize different allophonic variations in speech.

Applications

There are many practical benefits to be derived from the specification of the acoustic correlates of talker identity. If the reliable talker-specific characteristics of the speech signal were thoroughly understood, it would be possible to create sets of talker-to-sound rules similar in function to the letter-to-sound rules that are currently in use by text-to-speech systems. With letter-to-sound rules the acoustic specification of orthography and phonetic transcription is known in such detail that it is possible to produce speech based only on a sequence of alphanumeric characters. These rules allow the compact storage and transmission of speech information. However, only the linguistic content of the information is stored. Indexical properties such as talker identity, sex or age are lost. If talker identity could be specified by a set of rules, it would be possible to transmit or store the appropriate talker identity parameters along with the text so that two

of the most important characteristics of the speech waveform, the linguistic message and the talker's identity could be preserved and reproduced at will. The application of this knowledge would also be useful in speech recognition and in the forensic area as a means of identifying individuals by voice.

Finally, a thorough understanding of the acoustic correlates of talker identity would aid in the study of human perception and memory. If a talker's identity could be manipulated independently from the phonetic information, a separate and ecologically relevent dimension of the acoustic waveform could be specified in the construction of speech stimuli. Meaningful stimulus description is one of the major difficulties in experimental psychology, and with adequate knowledge of this indexical property of the speech signal, one more tool will be available for the study of the perception, coding, and retrieval of information by humans.

References

Abberton, E., & Fourcin, A. J. Intonation and speaker identification. Language and Speech, 1978, 21, 305-318.

Allard, F. and Henderson, L. Physical and name codes in auditory memory: The pursuit of an analogy. Quarterly Journal of Experimental Psychology, 1976, 28, 475-482.

Atal, B. S. Automatic speaker recognition based on pitch contours. Journal of the Acoustical Society of America, 1972, 52, 1687-1697.

Bernacki, R. JOT: Improved graphics capabilities for KLTEXC. In Research on Speech Perception Progress Report No. 8. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1982.

Carr, P. B. & Trill, D. Long-term larynx-exitation spectra. Journal of the Acoustical Society of America, 1964, 36, 2033-2040.

Carrell, T. D. & Kewley-Port D. K. Graphic support for KLTEXC. In Research on Speech Perception Progress Report No. 4. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1978.

Cole, R., Coltheart, M., & Allard, F. Memory of a speaker's voice: Reaction time to same- or different-voiced letters. Quarterly Journal of Experimental Psychology, 1974, 26, 1-7.

Coleman, R. O. Male and female voice quality and its relationship to vowel formant frequencies. Journal of Speech and Hearing Research, 1971, 14, 565-577.

Coleman, R. O. Speaker identification in the absence of inter- subject differences in glottal source characteristics. Journal of the Acoustical Society of America, 1973, 53, 1741-1743.

Coleman, R. O. A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. Journal of Speech and Hearing Research, 1976, 19, 168-180.

Cooper, F. S., Dellattre. P.C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 24, 597-606.

Craik, F. I. M. & Kirsner, K. The effect of speaker's voice on word recognition. Quarterly Journal of Experimental Psychology, 1974, 26, 274-284.

Drewnowski, A. & Murdock, B. B., Jr. The role of auditory features in memory span for words. Journal of Experimental Psychology: Human Learning and Memory, 1980, 6, 319-332.

Dunn, H. K. Methods of measuring vowel formant bandwidths. Journal of the Acoustical Society of America, 1961, 33, 1737-1746.

Fant, G. Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

Fant, G. A note on the vocal tract size factors and nonuniform F-pattern scalings. Speech Transmission Laboratory Quarterly Progress and Status Report No. 4, Stockholm, Sweden: Royal Institute of Technology, 1966.

Fant, G. Speech Sounds and Features. Cambridge, Mass: MIT Press, 1973.

Egan, J. P. Articulation and testing methods. Laryngoscope, 1948, 58, 955-991.

Flanagan, J. F. Some properties of the glottal sound source. Journal of Speech and Hearing Research, 1958, 1, 99-116.

Geiselman, R. E & Bellezza, F. S. Long-term memory for speaker's voice and source location. Memory and Cognition, 1976, 4, 483-489.

Glenn, J. W. and Kleiner N. Speaker identification based on nasal phonation. Journal of the Acoustical Society of America, 1968, 43, 368-372.

Goldstein, U. G. Speaker identifying features based on formant tracks. Journal of the Acoustical Society of America, 1976, 59, 176-182.

Hecker, M. H. L. Speaker recognition: An interpretive survey of the literature. ASHA Monographs, No. 16, 1971.

Hood, J. D. & Poole, J. P. Influence of the speaker and other factors affecting speech intelligibility. Audiology, 1980, 19, 434-455.

Hillman, R. E. & Weinberg, B. Estimation of the glottal volume velocity waveform properties: A review and study of some methodological assumptions. In N. Lass (Ed.) Speech and Language: Advances in Basic Research and Practice (Vol. 6). New York: Academic Press, 1981.

Ishizaka, K. & Flanagan, J. L. Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell System Technical Journal, 1962, 51, 1233-1268.

Ingemann, F. Identification of the speaker's sex from voiceless fricatives. Journal of the Acoustical Society of America, 1968, 44, 1142-1144.

Joos, M. A. Acoustic Phonetics. Language, 1948, Suppl. 24, 1-136.

Kewley-Port D. K. KLTEXC: Executive program to implement the Klatt software synthesizer. In Research on Speech Perception Progress Report No. 4. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1978.

Kewley-Port D. K. SPECTRUM: A program for analyzing the properties of speech. In Research on Speech Perception Progress Report No. 5. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1979.

Klatt, D. H. Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 1980, 67, 971-995.

Kramer, C. Perceptions of female and male speech. Language and Speech, 1977, 20, 151-161.

LaRiviere, C. Contributions of fundamental frequency and formant frequencies to speaker identification. Phonetica, 1975, 31, 185-1975.

Lashley, K. S. In search of the engram. In: Society for Experimental Biology (Great Britain) Physiological Mechanisms in Animal Behavior, 454-482. New York: Academic Press, 1950.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. Speaker sex identification from voiced, whispered, and filtered isolated vowels. Journal of the Acoustical Society of America, 1976, 59, 675-678.

Lehiste, I. & Meltzer, D. Vowel and speaker identification in natural and synthetic speech. Language and Speech, 1973, 16(4), 356-364.

Liberman, A. M., Cooper, F. S., Shankweiller, D.P., & Studdert-Kennedy, M. Perception of the Speech Code. Psychological Review, 1967, 74, 431-461.

Lieberman, P. Speech Physiology and Acoustic Phonetics. New York: Macmillan, 1977.

Luce, P. A. & Carrell, T. D. Creating and editing waveforms using WAVES. In Research on Speech Perception Progress Report No. 7. Bloomington, Indiana: Speech Research Laboratory, Indiana University, 1981.

Luce, P. A., Feustel, T.C., & Pisoni, D. B. Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 1983, 25, 17-32.

Markel J. D. & Gray, A. H., Jr. Linear Prediction of Speech. New York: Springer-Verlag, 1976.

Miller, J. E. Decapitation and recapitation, a study of voice quality. Paper presented at the Sixty-Eighth Meeting of the Acoustical Society of America, October 1964, Austin Texas.

Monsen, R. B. & Engebretson, A. M. Study of variations in the male and female glottal wave. Journal of the Acoustical Society of America, 1977, 62, 981-993.

Neter, J. & Wasserman, W. Applied linear statistical models. Homewood, Illinois: Richard Irwin, 1974.

Peterson, G. E. & Barney, H. L. Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 1952, 24, 175-184.

Pisoni, D. B. & Hunnicutt, S. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 572-575.

Pollack, I., Pickett, J. M., & Sumby, W. H. On the identification of speakers by voice. Journal of the Acoustical Society of America, 1954, 26, 403-406.

Posner, M. Abstraction and the process of recognition. In K. W. Spence and J. T. Spence (Eds.), The Psychology of Learning and Motivation, Volume 3. New York: Academic Press, 1969.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. Speech perception without traditional speech cues. Science, 1981, 212, 947-950.

Rosenberg, A. E. Effect of natural pulse shape on the quality of natural vowels. Journal of the Acoustical Society of America, 1971, 2, 583-590.

Sato, H. Acoustic cues of female voice quality. Electronics and Communication in Japan, 1974, 57-A, 29-38.

Schwartz, M. F. & Rine, H. E. Identification of speaker sex from isolated, whispered vowels. Journal of the Acoustical Society of America, 1968, 44, 1736-1737.

Sharf, D. J. Variations in vowel intensity measurements. Language and Speech, 1966, 9, 250-256.

Sondhi, M. M. Measurement of the glottal waveform. Journal of the Acoustical Society of America, 1975, 57, 228-232.

Stevens, K. N. Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. In A. Rigault & R. Charbonneau (Eds.), Proceedings of the 7th Internationial Congress of Phonetic Sciences, Montreal. The Hague: Mouton, 1972, 206-232.

Stevens, K. N., Kasowski, S., & Fant, G. An electrical analogy of the vocal tract. Journal of the Acoustical Society of America, 1953, 25, 734-742.

Wolf, J. J. Acoustic measurements for speaker recognition. Unpublished doctoral dissertation, M.I.T., Cambridge, 1969.

Wolf, J. J. Effecient acoustic parameters for speaker recognition. Journal of the Acoustical Society of America, 1972, 51, 2044-2056.

Appendix 1

Three sets of glottal waveforms were recorded from each of the six talkers on three different days as described in Chapters 2 and 4. A minimum of 24 hours intervened between each recording session. Talkers were instructed to produce high, medium, and low pitched vowels at each session. Each figure shows the three glottal waveforms produced by a single talker on each of the three days. These waveforms were not normalized in either amplitude or frequency for this display.

All the waveforms shown in this appendix were used as the test stimuli in Experiment 2, and the medium "pitch" waveforms were used to generate source functions for the synthetic stimuli used in Experiments 3 and 4.

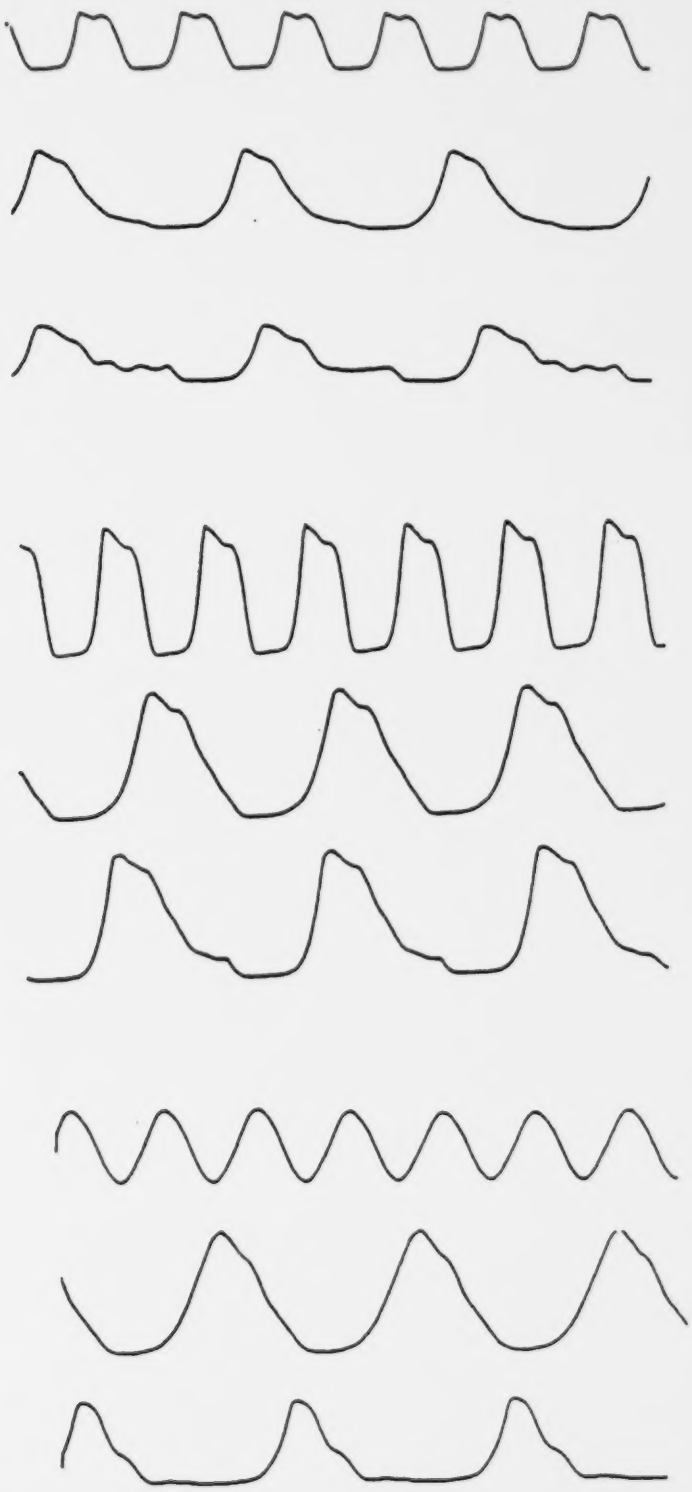


Figure A1.1. Glottal waveforms recorded from talker P.

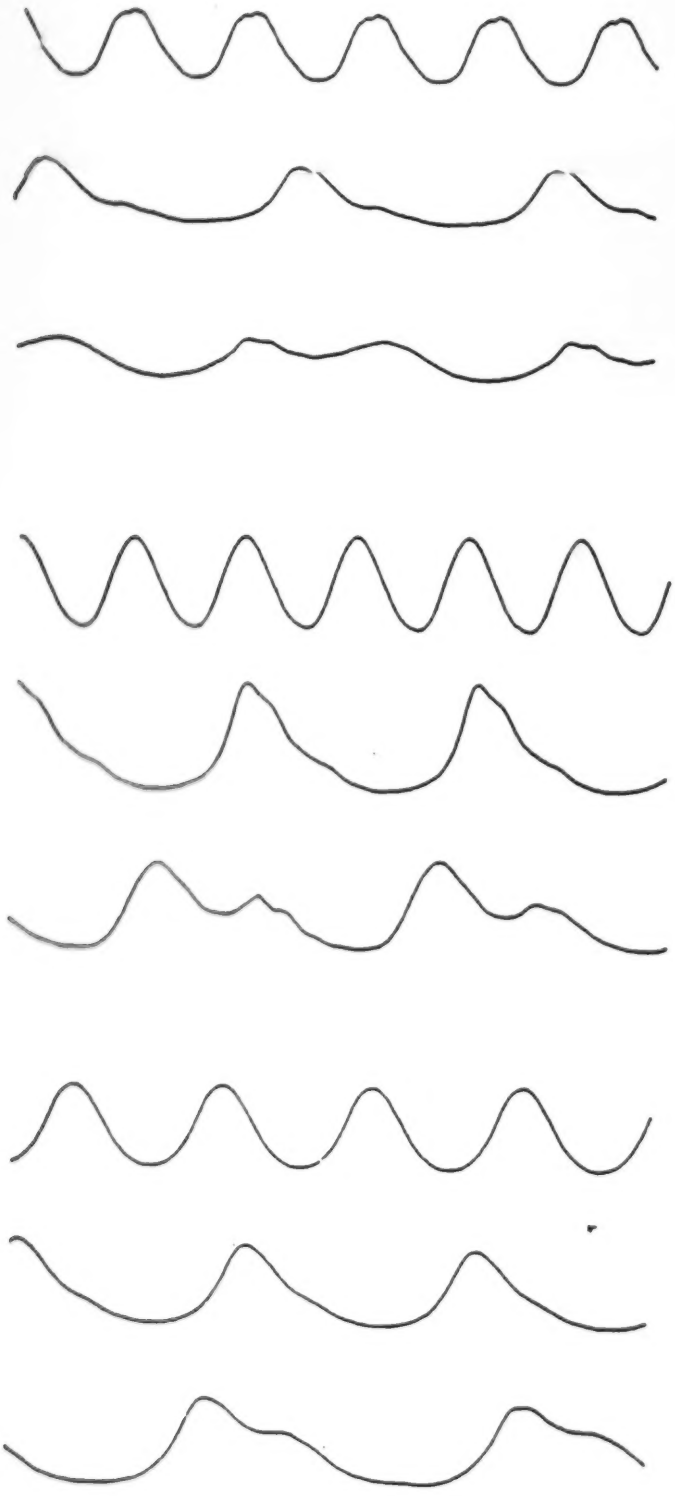


Figure A1.2. Glottal waveforms recorded from talker M.

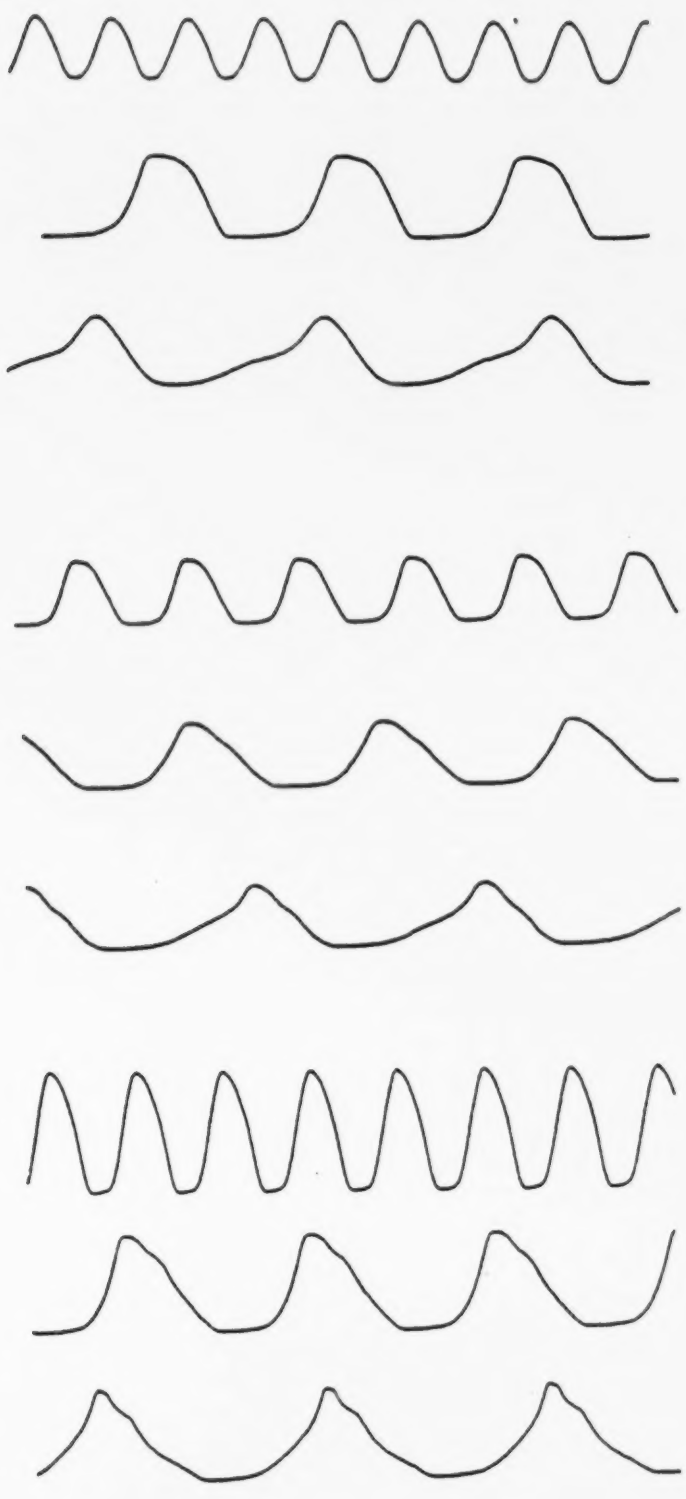


Figure A1.3. Glottal waveforms recorded from talker T.

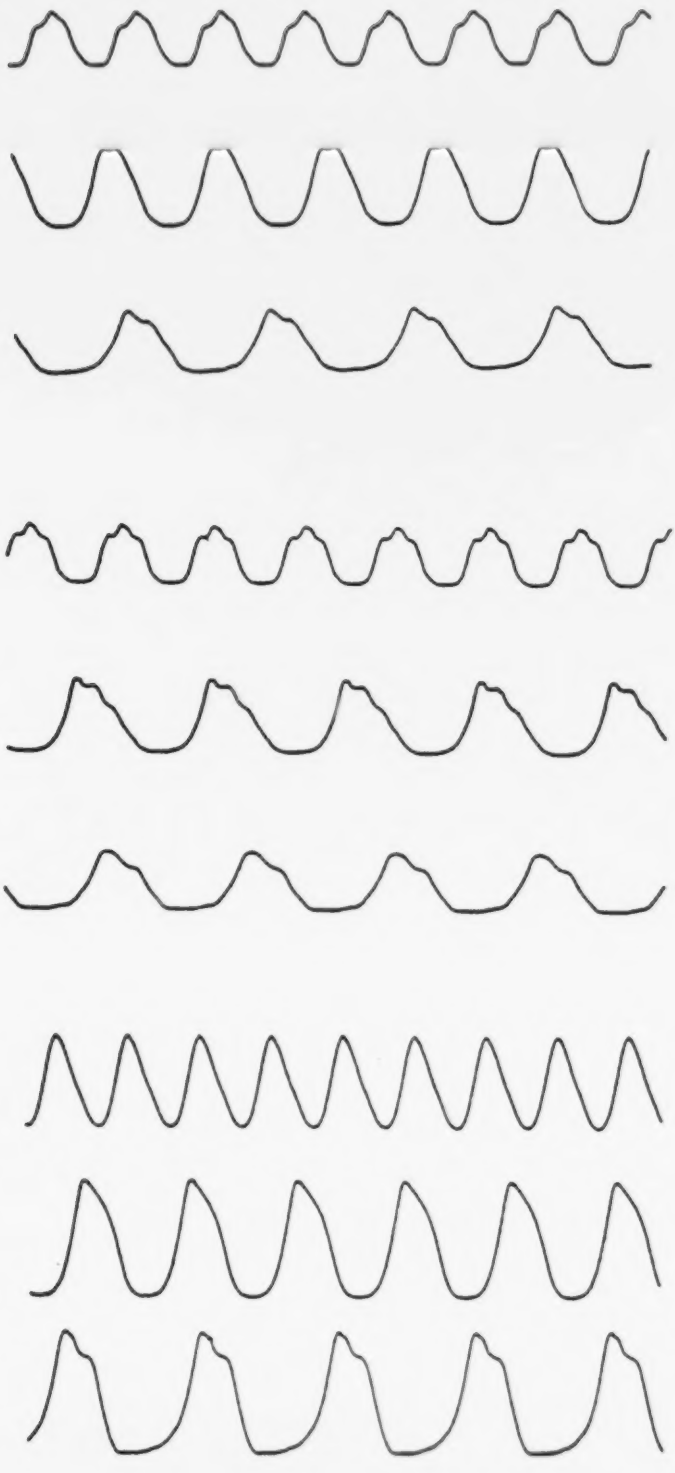


Figure A1.4. Glottal waveforms recorded from talker L.

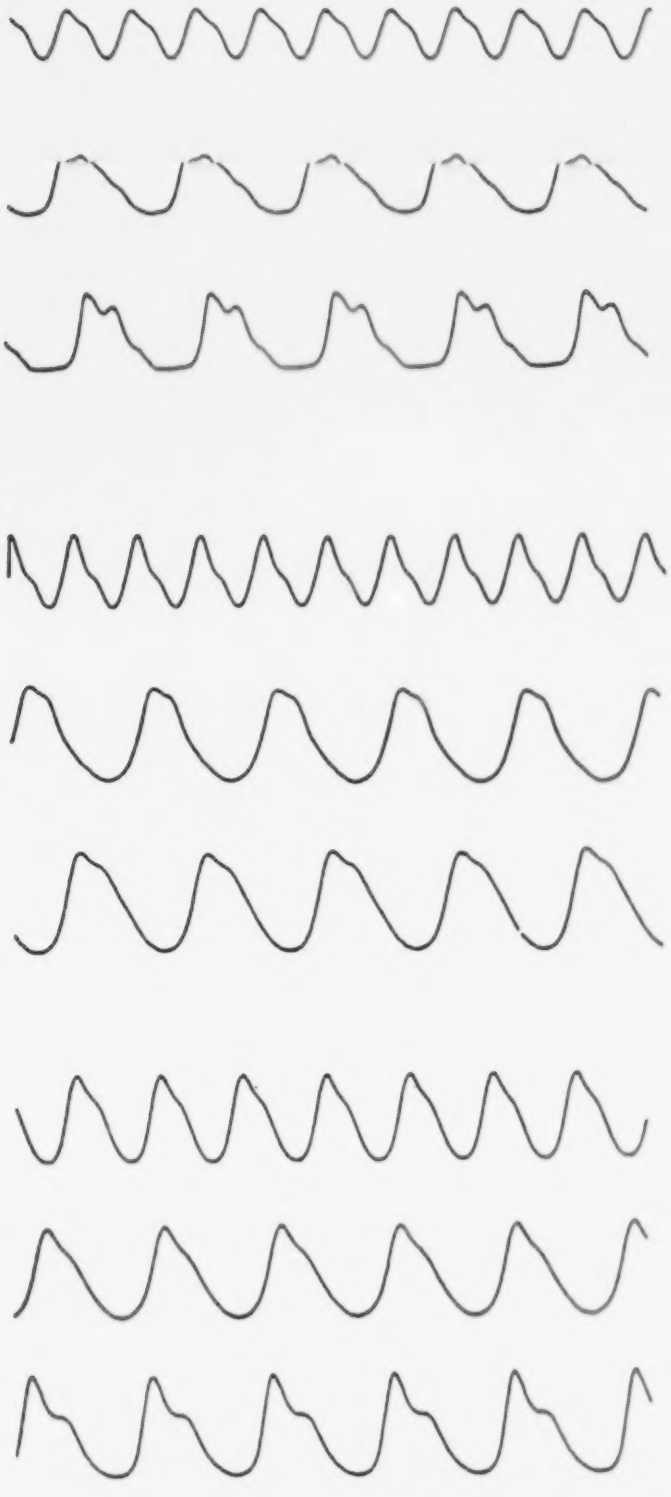


Figure A1.5. Glottal waveforms recorded from talker N.

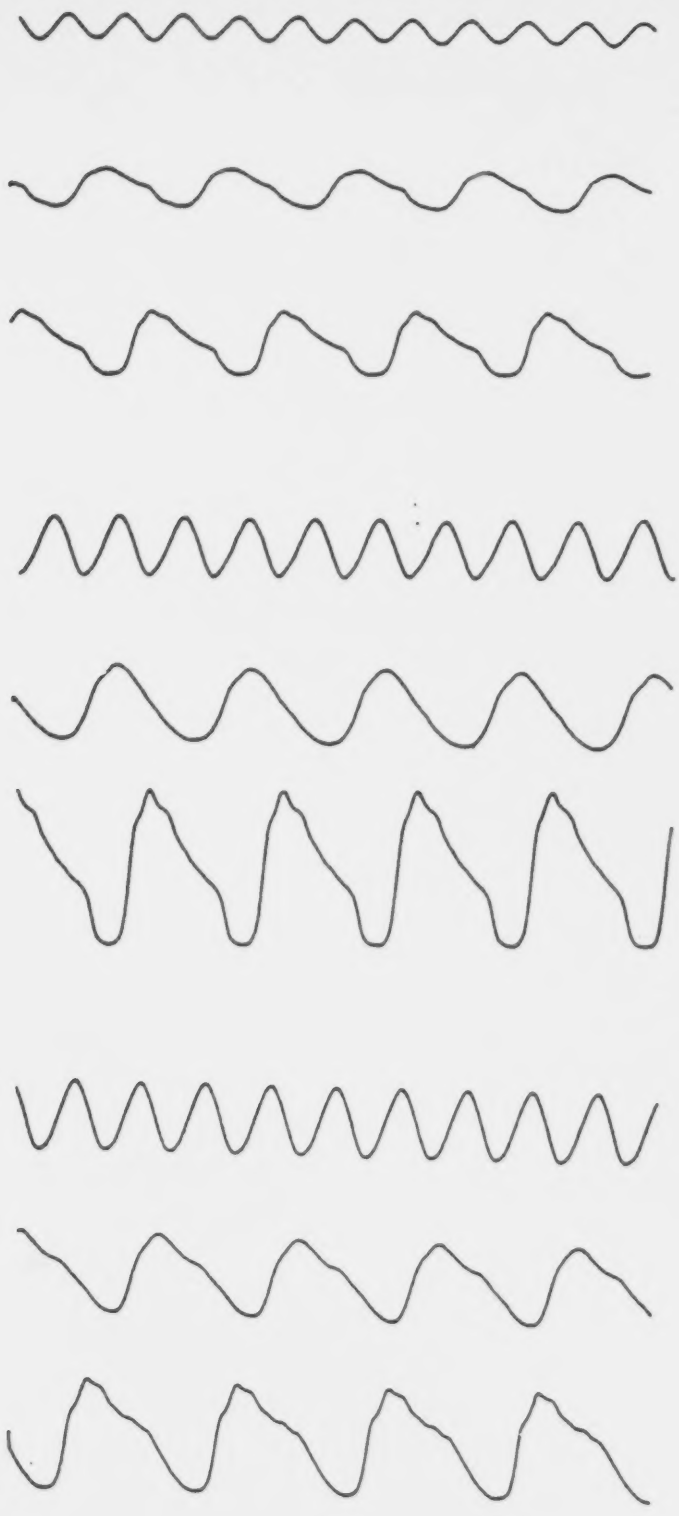


Figure A1.6. Glottal waveforms recorded from talker J.

Appendix 2

Fourier transforms were performed on the medium "pitch" glottal waveforms recorded from 6 talkers as described in Chapters 2 and 4. Each figure in Appendix 2 shows the amplitude spectra of these transforms for a single subject. The ordinate is the relative amplitude of each harmonic and the abscissa is the harmonic number. These amplitude spectra served as input to a discriminant analysis described in Chapter 4.

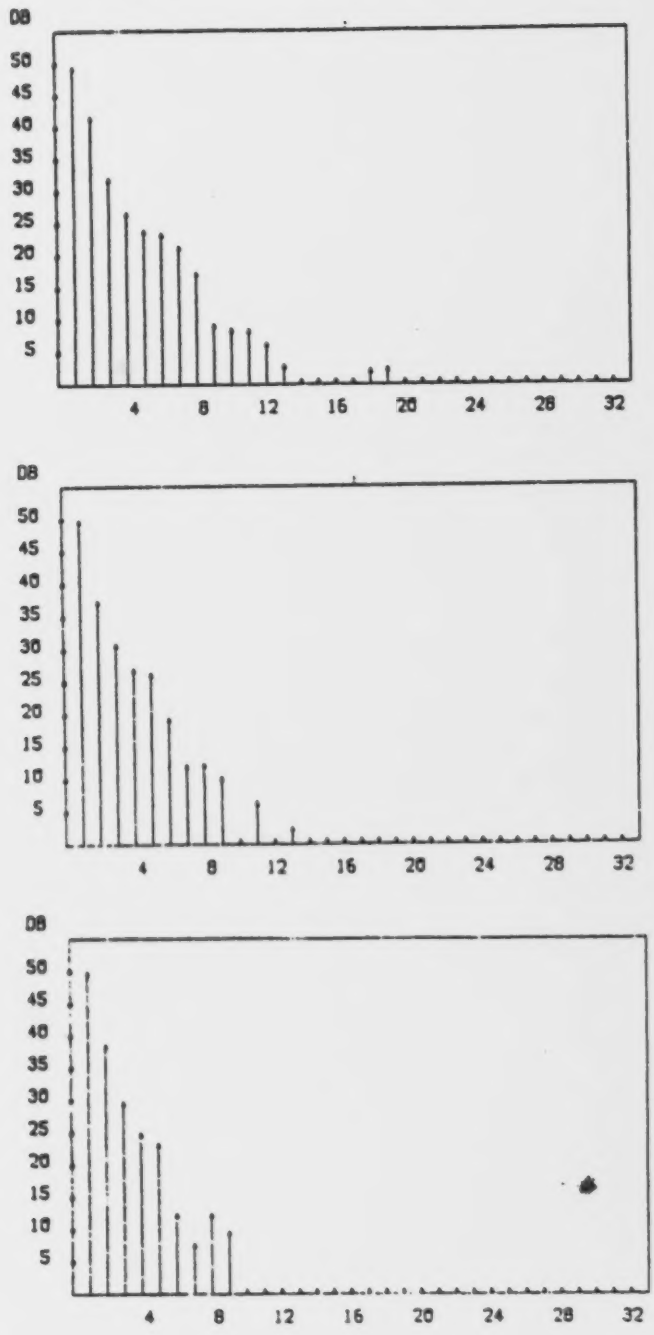


Figure A2.1. Fourier transforms of three glottal waveforms recorded from talker P.

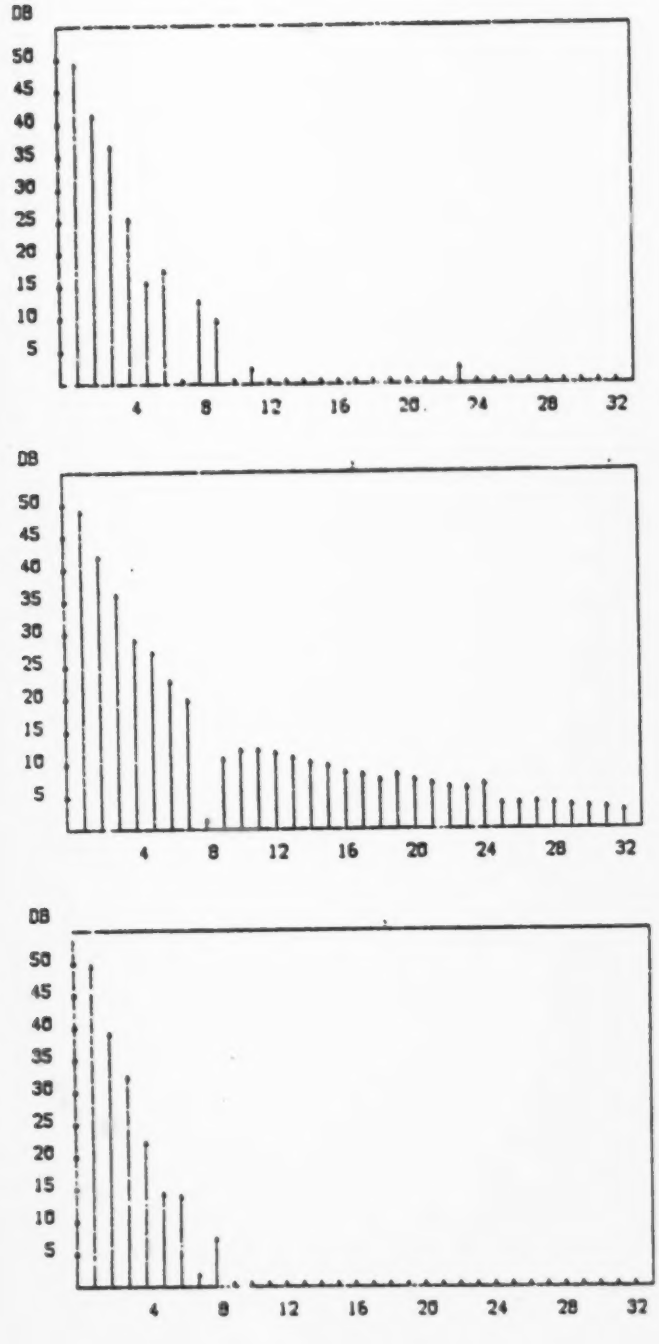


Figure A2.2. Fourier transforms of three glottal waveforms recorded from talker M.

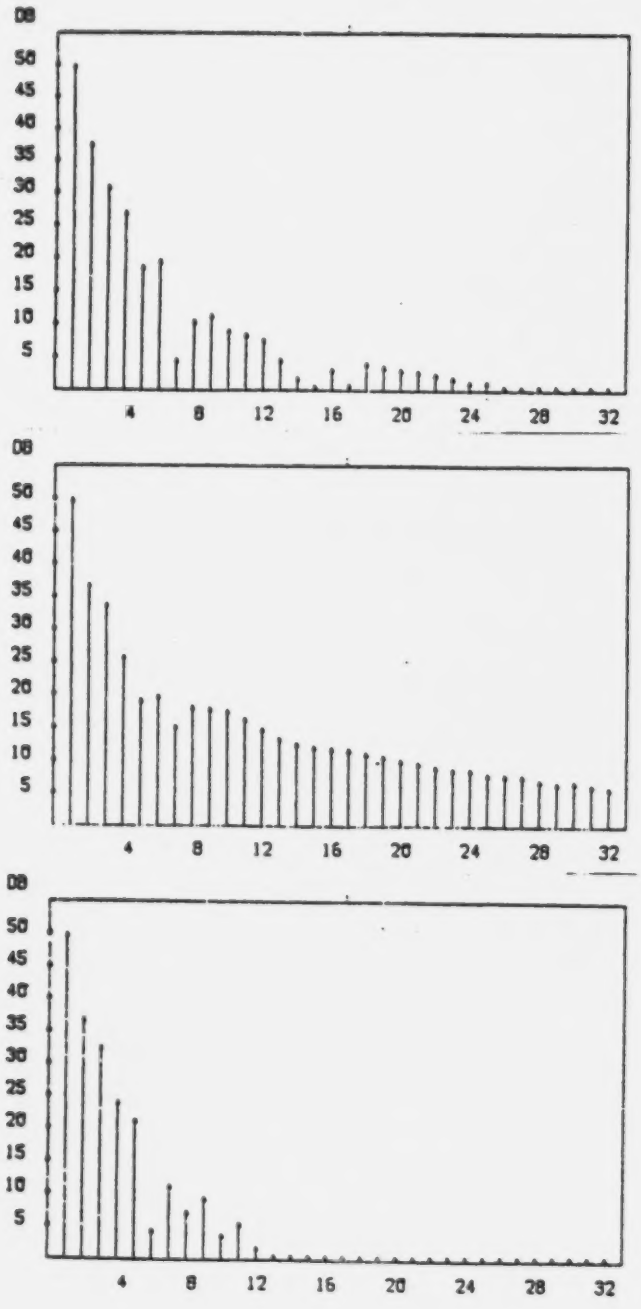


Figure A2.3. Fourier transforms of three glottal waveforms recorded from talker T.

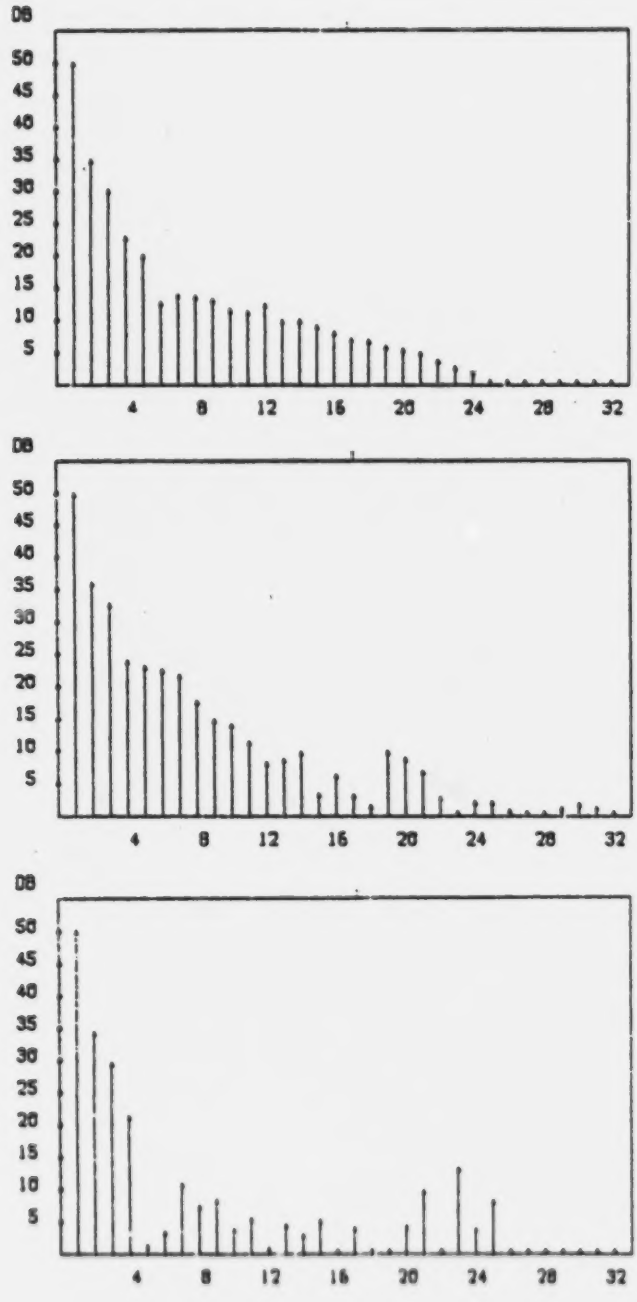


Figure A2.4. Fourier transforms of three glottal waveforms recorded from talker L.

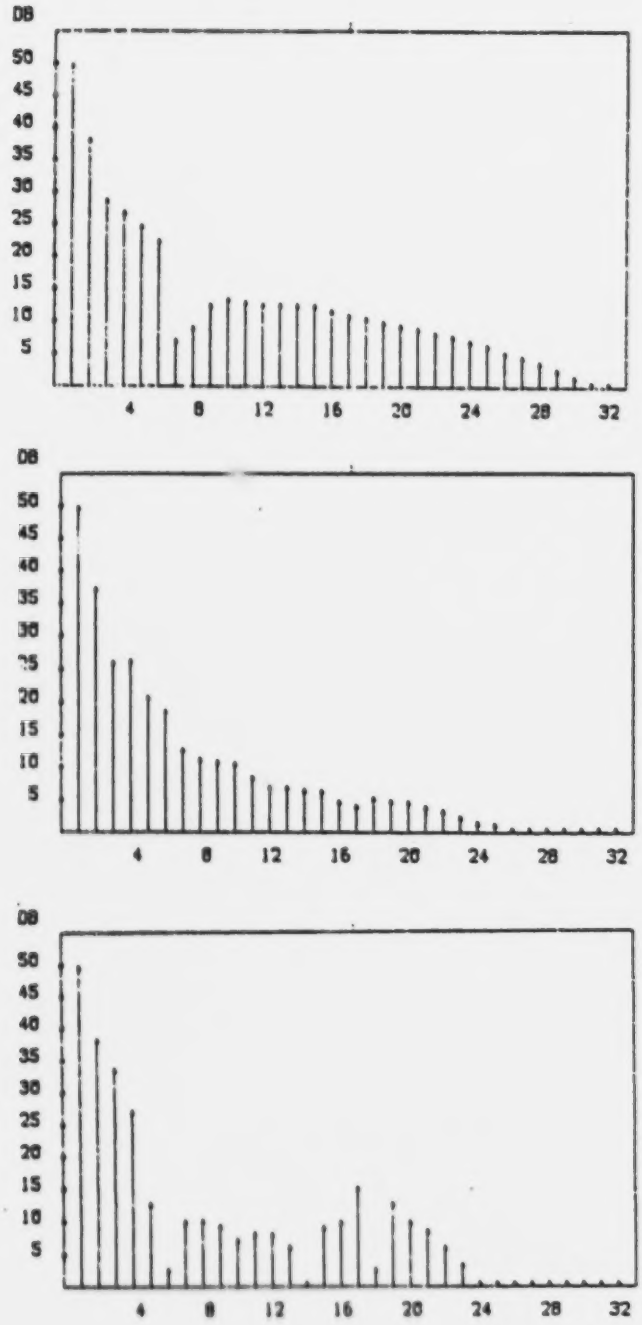


Figure A2.5. Fourier transforms of three glottal waveforms recorded from talker N.

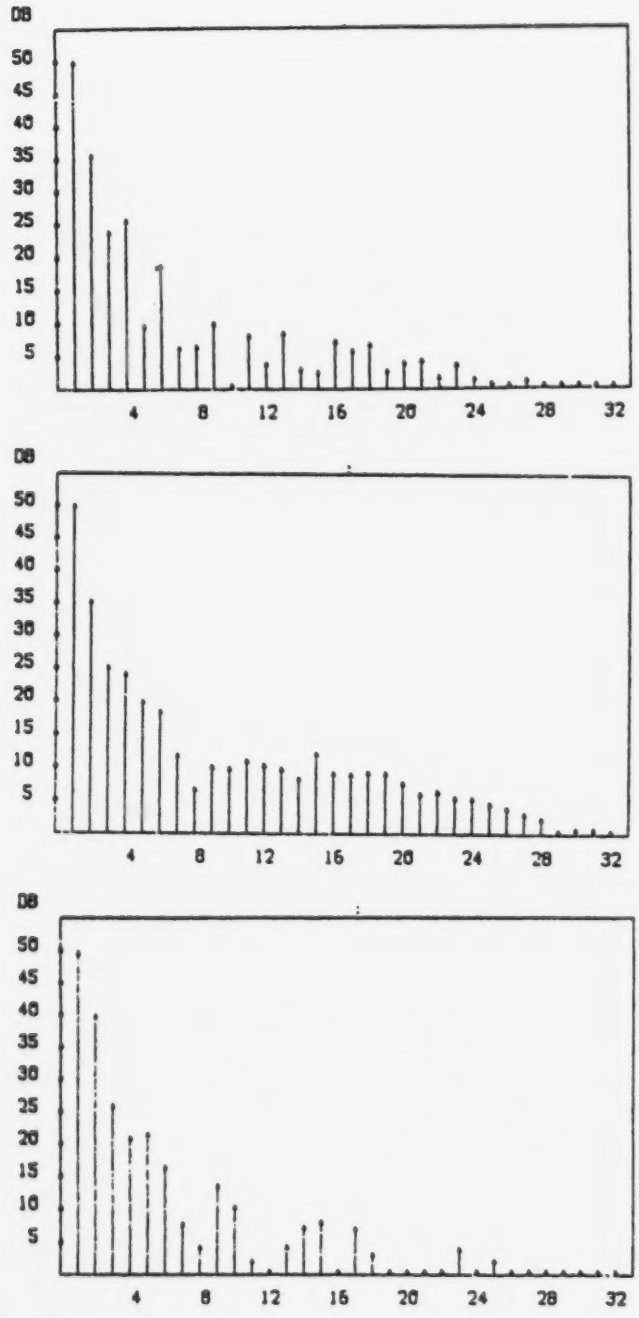


Figure A2.6. Fourier transforms of three glottal waveforms recorded from talker J.

Appendix 3

Table A3.1 shows listener's responses to each stimulus from the factorial stimulus set described in Chapter 6. The table is divided into four major columns and five major rows. Each column is further subdivided into four minor columns and each major row into four minor rows. Each of the major columns presents the proportion of responses listeners attributed to a particular talker under all stimulus conditions and each minor column shows the data from only those stimulus conditions synthesized with one particular glottal source. Each of the five major rows contains the response proportions of stimuli at a particular fundamental frequency and each of the minor rows specifies only the data from those stimulus conditions synthesized with a formant pattern based on one particular talker. For example, locate the number .319, the bottom left value in the block of data at the second major column and first major row. This value is the proportion of "T" responses for all stimuli that were synthesized with the glottal waveform of talker P, the formant pattern from talker N, and a fundamental frequency of 110 Hz. This mean proportion was taken from three repetitions of three different words, that shared the characteristics specified above, averaged across 13 listeners.

Table A3.1

Talker Response Probabilities for all Stimuli

		Response = P				Response = T				Response = L				Response = N			
		P	T	L	N	P	T	L	N	P	T	L	N	P	T	L	N
110 Hz	P	.504	.545	.508	.612	.299	.25	.280	.224	.111	.125	.127	.069	.085	.080	.085	.095
	T	.440	.545	.496	.500	.388	.268	.256	.284	.121	.107	.145	.147	.052	.080	.103	.069
	L	.360	.461	.371	.363	.279	.278	.181	.310	.279	.148	.353	.239	.081	.113	.095	.088
	N	.353	.425	.448	.474	.319	.257	.259	.310	.259	.265	.190	.138	.069	.053	.103	.078
140 Hz	P	.360	.579	.461	.383	.430	.263	.304	.374	.061	.088	.122	.157	.149	.070	.113	.087
	T	.209	.395	.368	.316	.617	.404	.421	.513	.087	.149	.123	.085	.087	.053	.088	.085
	L	.248	.314	.278	.171	.274	.254	.183	.231	.359	.305	.383	.496	.120	.127	.157	.103
	N	.233	.376	.359	.295	.405	.291	.222	.375	.241	.248	.308	.223	.121	.085	.111	.107
170 Hz	P	.175	.272	.286	.265	.614	.386	.387	.462	.114	.211	.210	.179	.096	.132	.118	.094
	T	.177	.190	.172	.242	.593	.621	.672	.55	.150	.121	.095	.083	.080	.069	.060	.125
	L	.138	.105	.123	.138	.198	.088	.149	.138	.509	.667	.649	.621	.133	.140	.079	.103
	N	.144	.177	.186	.183	.279	.230	.178	.257	.342	.381	.475	.339	.234	.212	.161	.220
200 Hz	P	.142	.138	.114	.154	.327	.207	.263	.299	.319	.362	.342	.316	.212	.293	.281	.231
	T	.221	.117	.127	.164	.407	.441	.418	.414	.248	.207	.200	.233	.124	.234	.255	.190
	L	.093	.093	.142	.121	.136	.161	.183	.147	.542	.610	.542	.569	.229	.136	.133	.164
	N	.214	.081	.096	.142	.145	.135	.114	.195	.376	.514	.544	.434	.265	.270	.246	.230
230 Hz	P	.150	.173	.153	.088	.195	.173	.225	.186	.212	.291	.297	.292	.442	.364	.324	.434
	T	.070	.036	.183	.149	.296	.291	.296	.289	.217	.300	.226	.246	.417	.373	.296	.316
	L	.102	.130	.085	.122	.127	.087	.085	.096	.449	.504	.547	.548	.322	.278	.282	.235
	N	.085	.167	.122	.144	.195	.149	.148	.093	.263	.307	.374	.364	.458	.377	.357	.398

Note. Single column labels specify the glottal source.

Single row labels specify the formant source.